

1998

Spatial prediction, spatial sampling, and measurement error

Walter Jeremy Koch Aldworth
Iowa State University

Follow this and additional works at: <https://lib.dr.iastate.edu/rtd>



Part of the [Statistics and Probability Commons](#)

Recommended Citation

Aldworth, Walter Jeremy Koch, "Spatial prediction, spatial sampling, and measurement error " (1998). *Retrospective Theses and Dissertations*. 11842.
<https://lib.dr.iastate.edu/rtd/11842>

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Spatial prediction, spatial sampling, and measurement error

by

Walter Jeremy Koch Aldworth

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Major Professor: Noel A. C. Cressie

Iowa State University

Ames, Iowa

1998

Copyright © Walter Jeremy Koch Aldworth, 1998. All rights reserved.

UMI Number: 9841034

**Copyright 1998 by
Aldworth, Walter Jeremy Koch**

All rights reserved.

**UMI Microform 9841034
Copyright 1998, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of
Walter Jeremy Koch Aldworth
has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

DEDICATION

To Sue, who always believed I could do it.

Come, numbers, be like seasons all,
signs of my years, perpetual,
waves of the day upon the white
wide ledger, shore of ink and sight.
The world in columns asks me why
I cannot balance sea and sky.

—Sydney Clouts

To see a World in a Grain of Sand
And a Heaven in a Wild Flower,
Hold infinity in the palm of your hand
And Eternity in an hour.

—William Blake

TABLE OF CONTENTS

GENERAL INTRODUCTION	1
1 Dissertation Organization	3
References	4
SAMPLING DESIGNS AND PREDICTION METHODS FOR GAUSSIAN SPATIAL PROCESSES	6
1 Introduction	7
2 Brief Review of Geostatistical Theory, Survey-Sampling Theory, and Spatial Sampling	8
2.1 Geostatistical Theory	8
2.1.1 The Spatial Process	8
2.1.2 Stationarity	9
2.1.3 Ordinary Kriging	9
2.1.4 Constrained Kriging	12
2.1.5 Trend	14
2.1.6 Measurement Error	14
2.1.7 Estimating and Modeling the Variogram	15
2.2 Survey-Sampling Theory	16
2.2.1 Finite-Population Sampling	17
2.2.2 Horvitz-Thompson Estimation for a Continuous Population . .	20
2.2.3 Superpopulation Models in Survey Sampling	21
2.2.4 Measurement Error	22
2.2.5 Estimation in the Presence of Measurement Error	22
2.2.6 Inference in Spatial Sampling	25

2.3	Review of the Spatial Sampling Literature	26
2.3.1	Model-Based and Design-Based Sampling Approaches	26
2.3.2	Design Criteria	27
2.3.3	Comparison of Sampling Designs	32
2.3.4	Adaptive Sampling	35
3	Computer Simulation Experiment	35
3.1	Gaussian Spatial Models	36
3.2	Factors of the Computer Experiment	37
3.2.1	Regions	37
3.2.2	Strength of Spatial Dependence	37
3.2.3	Trend	37
3.2.4	Noise	37
3.2.5	Realizations of the State Process	38
3.2.6	Sampling Designs	38
3.2.7	Prediction Methods	40
3.3	Responses of the Computer Simulation Experiment	45
3.3.1	Spatial-Mean Responses	46
3.3.2	SCDF Responses	47
3.4	Comments	47
4	Results of the Experiment	48
4.1	Spatial-Mean Prediction Over the Local Region	49
4.2	SCDF Prediction Over the Global Region	53
5	Conclusions	58
	Appendix	59
	References	61
	Acknowledgements	64
	THE PREDICTION OF MEASURABLE FUNCTIONALS OF SPATIAL PROCESSES USING DATA CONTAMINATED WITH MEASUREMENT ERROR	65

1	Introduction	65
2	Measurement-error Spatial Model	68
3	Prediction of Linear Functionals of the S -process	71
3.1	Best Predictor	73
3.2	Linear Predictors	73
3.2.1	Best (Heterogeneous) Linear Predictor	73
3.2.2	Best Linear Unbiased Predictor	74
3.3	Estimation of $\{\theta, \tau^2\}$	76
3.4	Interpolation and Measurement Error	78
3.5	Kriging and Uncorrelated Spatial Processes	79
3.6	Multivariate Prediction	82
3.6.1	Cokriging	83
3.6.2	Generalized Universal Kriging	86
3.6.3	Generalized Cokriging	88
3.7	Nonlinear Predictors	88
3.7.1	Indicator Kriging	89
3.7.2	Indicator Cokriging	90
3.7.3	Disjunctive Kriging	91
4	Prediction of Nonlinear Functionals of the S -process	92
4.1	Best Predictor	93
4.2	Constrained Kriging	93
4.2.1	Unbiasedness of $g(\mathbf{a}'\mathbf{Z})$	94
4.2.2	Optimization of $g(\mathbf{a}'\mathbf{Z})$	95
4.2.3	Constrained Kriging Predictors	95
4.2.4	Existence of the Constrained Kriging Predictor	97
4.2.5	Geometric Interpretation of Constrained Kriging	99
4.2.6	Affine Correction	105
4.3	Constrained Kriging and Uncorrelated Spatial Processes	109

4.4	Constrained Kriging and Finite-population Sampling	111
4.4.1	Finite-population Sampling	112
4.4.2	Horvitz-Thompson Estimation	113
4.4.3	Measurement Error and HTE Bias	115
4.4.4	Alternative CDF estimators	116
4.5	Constrained Kriging and Conditional Simulation	121
4.5.1	Orthogonality Properties of Simple Kriging and Ordinary Kriging	121
4.5.2	Conditional Simulation and Measurement Error	122
4.5.3	Conditional Simulation and Constrained Kriging	127
5	Covariance-Matching Constrained Kriging	130
5.1	Covariance-Matching Constrained Kriging Equations	132
5.1.1	Unbiasedness of $g(\mathbf{A}'\mathbf{Z})$	132
5.1.2	Optimization of $g(\mathbf{A}'\mathbf{Z})$	133
5.1.3	Covariance-Matching Constrained Kriging Predictor	141
5.1.4	Existence of the Covariance-Matching Constrained Kriging Pre- dictor	143
5.2	Summary	145
6	Computer-Simulation Experiment	146
6.1	Spatial Domain and Spatial Model	146
6.2	Factors of the Experiment	147
6.3	Predictands	151
6.4	Responses of the Experiment	152
6.5	Results of the Experiment	152
6.6	Summary	170
7	Spatial Analysis of Particulate Matter in Pittsburgh Area	170
7.1	Exploratory Data Analysis	170
7.2	Modeling the Semivariogram	171
7.3	Spatial Prediction of PM_{10}	176
8	Conclusions	180

References	181
Acknowledgements	183
GENERAL CONCLUSIONS	184
ACKNOWLEDGEMENTS	186

GENERAL INTRODUCTION

A *spatial process* is a real- (or vector-) valued stochastic process

$$Z(\cdot) \equiv \{Z(s) : s \in D\},$$

where $D \subset \mathbb{R}^d$. In a geostatistical context, D is usually assumed to be fixed set of positive d -dimensional volume. Sometimes it is useful to work with a discretized version of D , that is, a finite collection of points arranged in a grid over D . In this dissertation, $Z(\cdot)$ is assumed to be an observable process representing actual and potential data.

The thread that binds this dissertation together is the explicit acknowledgement that data are almost always measured with error, and this is expressed in the *measurement-error spatial model*, defined as

$$Z(s) = S(s) + \epsilon(s); \quad s \in D,$$

where $S(\cdot)$ represents the “signal” process, and $\epsilon(\cdot)$ represents the measurement-error or “noise” process that usually accompanies any attempt to observe $S(\cdot)$. Both $S(\cdot)$ and $\epsilon(\cdot)$ are assumed to be unobservable processes, independent of each other.

The (actual) *data* from the spatial process $Z(\cdot)$ consist of the observations

$$\{Z(s_1), \dots, Z(s_n)\},$$

taken at locations in the *sample*, denoted as

$$A \equiv \{s_1, \dots, s_n\} \subset D.$$

In a geostatistical context, A is usually assumed to be fixed, although Aldworth and Cressie (1998) show how randomness in A may be considered within a geostatistical context.

For a phenomenon that varies over a continuous (or even a large finite) spatial domain, it is seldom feasible, or even possible, to observe every potential datum of some study variable

associated with that phenomenon. Thus, an important part of statistics is statistical sampling theory, where inference about the study variable may be made from a subset, or sample, of the population of potential data.

Spatial sampling refers to the sampling of geo-referenced or spatially labeled phenomena. In the spatial context, interest is usually in the *prediction* of (some function of) the study variable at multiple unsampled sites and it is in this sense that the prediction problem is multivariate. Given some predictand together with its predictor, a *best* sampling plan or network refers to the choice of locations at which to sample the phenomenon in order to achieve optimality according to a given criterion (e.g., minimize average mean squared prediction error, where the average is taken over multiple prediction locations). In practice, optimal sampling plans may be extremely difficult to achieve but good, although suboptimal, sampling plans may be relatively easy to obtain and these designs, at least, should be sought.

A commonly chosen predictand in survey sampling is the total (or mean) of the study variable over a specified spatial domain. In this dissertation, we shall also consider predictands defined over some “local” subregion of the domain, and predictands that are nonlinear functions of the study variable at multiple spatial locations. An extensive review of the spatial-sampling literature is given in this dissertation.

In much of the geostatistics literature, interest usually involves the prediction of some functional of $Z(\cdot)$ (e.g., Journel and Huijbregts, 1978; Rivoirard, 1994), thereby implicitly assuming that observations are exact measurements of the spatial process of interest. Very little in the geostatistics literature deals with a measurement-error component explicitly included in a spatial model. A measurement-error spatial model is first discussed by Cressie (1986) and further developed by Cressie (1988).

In addition, much attention has been given to the prediction of *linear* functionals of $S(\cdot)$ (i.e., kriging), such as prediction at some unsampled location or the prediction of the average over some subregion in the domain (e.g., Journel and Huijbregts, 1978; Cressie, 1993a). It has been shown by, for example, Gotway and Cressie (1993), that the ordinary-kriging and universal-kriging prediction methodology has useful optimality properties for predictands of

this kind, particularly if the distributional characteristics of the study variable (as seen in the data) appear to be Gaussian. Ordinary-kriging and universal-kriging predictors are linear in the data and are simple to construct (e.g., Journel and Huijbregts, 1978, Ch. V; Cressie, 1993a, Ch. 3), and this prediction methodology can be extended easily to the case where the data are multivariate (e.g., Myers, 1982; Ver Hoef and Cressie, 1993). However, if the data suggest that a nonGaussian distribution is appropriate for the study variable, then these linear kriging predictors may be far from optimal. In such a case, alternative predictors, nonlinear in the data, such as indicator kriging, indicator cokriging, and disjunctive kriging, have been proposed as alternatives to the linear kriging predictors (e.g., Journel and Huijbregts, 1978, Ch. VIII; Cressie, 1993a, Sect. 5.1).

The problem of the prediction of *nonlinear* functionals of $S(\cdot)$ (Gaussian or not) is more difficult. For such problems, the linear kriging predictors (i.e., ordinary and universal kriging) generally perform poorly because they are usually too smooth (e.g., Journel and Huijbregts, 1978, Ch. VIII). Nonlinear kriging predictors (i.e., indicator kriging, indicator cokriging, and disjunctive kriging) may perform better than the linear kriging predictors but possibly only if the measurement error is not substantial; versions of them that account for measurement error have yet to be developed. An alternative predictor proposed by Cressie (1993b), is based on linear kriging methodology, but with a constraint forcing the variance of the predictor to match that of the predictand. This predictor, called “constrained kriging”, has useful unbiasedness properties, approximate second-order optimality properties, and it easily filters out additive measurement error. The theory supporting constrained kriging is extended to the “covariance-matching” case where multiple predictions are required simultaneously, and the resulting predictor is shown to have certain optimality properties.

1 Dissertation Organization

This dissertation consists of two distinct papers, preceded by a general introduction and followed by general concluding statements.

The first paper in this dissertation has been accepted for publication as a book chapter in

Multivariate Analysis, Design of Experiments and Survey Sampling, (1998), S. Ghosh (ed.), Marcel and Dekker, New York. In this paper, a brief review of geostatistical theory and survey-sampling theory, and an extensive review of the spatial-sampling literature are given in Section 2. Based on this knowledge, we designed a simulation experiment whose details are described in Section 3. Section 4 analyzes the results of the experiment and conclusions are given in Section 5.

The second paper is a longer version of an article that we wish to present for publication in a statistical journal. In this paper, a spatial model that explicitly includes a measurement-error component is presented in Section 2, and the effects of additive measurement error on kriging predictors, linear and nonlinear, are discussed in Section 3. Constrained kriging is discussed in Section 4, covariance-matching constrained kriging is presented in Section 5, and a simulation study comparing the different linear kriging predictors is presented Section 6. In Section 7 a study of particulate-matter data in the Pittsburgh area is presented where these predictors are applied. Finally, Section 8 contains discussion and conclusions.

References

- Aldworth, J. and Cressie, N. (1998), "Sampling designs and prediction methods for gaussian spatial processes", in *Multivariate Analysis, Design of Experiments and Survey Sampling*, S. Ghosh (ed.), Marcel Decker, New York, forthcoming.
- Cressie, N. (1986), "Kriging nonstationary data", *Journal of the American Statistical Association*, **81**, 625–634.
- Cressie, N. (1988), "Spatial prediction and ordinary kriging", *Mathematical Geology*, **20**, 405–421 [Erratum, *Mathematical Geology*, **21**, 493–494.].
- Cressie, N. (1993a), *Statistics for Spatial Data*, rev. ed., Wiley, New York.
- Cressie, N. (1993b), "Aggregation in geostatistical problems", in *Geostatistics Tróia 1992*, Vol. 1, A. Soares (ed.), Kluwer Academic Publishers, Dordrecht, pp. 25–36.
- Gotway, C. A. and Cressie, N. (1993), "Improved multivariate prediction under a general linear model", *Journal of Multivariate Analysis*, **45**, 56–72.

- Journel, A. G. and Huijbregts, C. J. (1978), *Mining Geostatistics*, Academic Press, London.
- Myers, D. E. (1982), "Matrix formulation of co-kriging", *Journal of the International Association of Mathematical Geology*, 14, 249–257.
- Rivoirard, J. (1994), *Introduction to Disjunctive Kriging and Non-linear Geostatistics*, Oxford University Press, Oxford.
- Ver Hoef, J. M. and Cressie, N. (1993), "Multivariable spatial prediction", *Mathematical Geology*, 25, 219–240 [Errata, *Mathematical Geology*, 26, 273–275.].

SAMPLING DESIGNS AND PREDICTION METHODS FOR GAUSSIAN SPATIAL PROCESSES

A paper accepted as a book chapter in *Multivariate Analysis, Design of Experiments and
Survey Sampling*

Jeremy Aldworth and Noel Cressie

Abstract

A geostatistical model can provide a powerful way of predicting unknown parts of some spatial phenomenon. The prediction problem is multivariate in the sense that one wishes to predict at multiple spatial locations. The results presented in this chapter offer compelling evidence that a geostatistical model should be incorporated into spatial sampling and analysis, where possible. Even when the observable process is contaminated with measurement error, there is a straightforward way to filter it out by appropriately modifying the spatial prediction equations. Our results show that a geostatistical analysis of a certain class of non-clustered designs, whether simple-random, stratified-random, or systematic-with-a-random-start, performs extremely well with respect to design-based optimality criteria. In contrast, clustered designs, corresponding to repeated sampling from “representative sites”, perform very poorly. One important aspect of our study is the prediction of spatial statistics defined over small areas (called local regions), that are subsets of a global region over which a network of sampling sites is chosen. Under circumstances where both local and nonlinear functions of the process are to be predicted, it is demonstrated that appropriate geostatistical analyses perform very well, irrespective of the (non-clustered) sampling design.

1 Introduction

For a phenomenon that varies over a continuous (or even a large finite) spatial domain, it is seldom feasible, or even possible, to observe every potential datum of some study variable associated with that phenomenon. Thus, an important part of statistics is statistical sampling theory, where inference about the study variable may be made from a subset, or sample, of the population of potential data.

Spatial sampling refers to the sampling of geo-referenced or spatially labeled phenomena. In the spatial context, interest is usually in the *prediction* of (some function of) the study variable at multiple unsampled sites and it is in this sense that the prediction problem is multivariate. Given some predictand together with its predictor, a *best* sampling plan or network refers to the choice of locations at which to sample the phenomenon in order to achieve optimality according to a given criterion (e.g., minimize average mean squared prediction error, where the average is taken over multiple prediction locations). In practice, optimal sampling plans may be extremely difficult to achieve but good, although suboptimal, sampling plans may be relatively easy to obtain and these designs, at least, should be sought.

A commonly chosen predictand in survey sampling is the total (or mean) of the study variable over a specified spatial domain. In this article, we shall also consider predictands defined over some “local” subregion of the domain, and predictands that are nonlinear functions of the study variable at multiple spatial locations.

The objective of this paper is to gauge, through a carefully designed simulation experiment, the performance of different prediction methods under different sampling designs, over several realizations of a spatial process whose strength of spatial dependence varies from zero to very strong. Included are both “spatial” and “non-spatial” analyses and designs. Our emphasis is on prediction of spatial statistics defined on both “local” and “global” regions, based on data obtained from a global network of sampling sites. A brief review of geostatistical theory, survey-sampling theory, and of the spatial-sampling literature are given in Section 2. Based on this knowledge, we designed a simulation experiment whose details are described in Section 3. Section 4 analyzes the results of the experiment and conclusions are given in Section 5.

2 Brief Review of Geostatistical Theory, Survey-Sampling Theory, and Spatial Sampling

2.1 Geostatistical Theory

Suppose some phenomenon of interest is indexed by spatial location in a domain $D \subset \mathbb{R}^2$. We wish to choose a sample size n and sample locations $\{s_1, \dots, s_n\} \subset D$ so that “good” inferences may be made about the phenomenon from the sample data. Such spatially labeled data often exhibit dependence in the sense that observations closer together tend to be more similar than observations farther apart, which should be exploited in the search for an optimal (or good) network of sites. A brief synopsis of the geostatistical theory characterizing this spatial dependence follows (see Cressie, 1993a, part I, for more details).

2.1.1 The Spatial Process

A spatial planar process is a real- (or vector-) valued stochastic process $\{Z(s) : s \in D\}$ where $D \subset \mathbb{R}^2$. In all that is to follow, the case of real-valued $Z(\cdot)$ will be considered; inference is desired on unobserved parts of the process at multiple locations.

In studies where spatially labeled data exhibit spatial dependence, the following model is useful:

$$Z(s) = \mu(s) + \delta(s); \quad s \in D, \quad (2.1)$$

where $\mu(\cdot)$ is the large-scale, deterministic, mean structure of the process (i.e., trend) and $\delta(\cdot)$ is the small-scale stochastic structure that models the spatial dependence among the data. That is,

$$E[\delta(s)] = 0; \quad s \in D \quad (2.2)$$

$$\text{cov}[\delta(s), \delta(u)] \equiv C(s, u); \quad s, u \in D. \quad (2.3)$$

Hence $E[Z(s)] = \mu(s); \quad s \in D$, and $\text{cov}[Z(s), Z(u)] = C(s, u); \quad s, u \in D$.

Another useful measure of spatial dependence is the *variogram*:

$$2\gamma(s, u) \equiv \text{var}[Z(s) - Z(u)]$$

$$= C(\mathbf{s}, \mathbf{s}) + C(\mathbf{u}, \mathbf{u}) - 2C(\mathbf{s}, \mathbf{u}); \quad \mathbf{s}, \mathbf{u} \in D. \quad (2.4)$$

The quantity $\gamma(\cdot, \cdot)$ is called the *semivariogram*.

2.1.2 Stationarity

The process $Z(\cdot)$ is *first-order stationary* if the following condition holds:

$$E[Z(\mathbf{s})] = \mu; \quad \mathbf{s} \in D. \quad (2.5)$$

If $Z(\cdot)$ is first-order stationary and if it satisfies

$$\text{var}[Z(\mathbf{s}) - Z(\mathbf{u})] \equiv 2\gamma(\mathbf{s}, \mathbf{u}) = 2\gamma^o(\mathbf{s} - \mathbf{u}); \quad \mathbf{s}, \mathbf{u} \in D, \quad (2.6)$$

then $Z(\cdot)$ is said to be *intrinsically stationary*. Note that the variogram $2\gamma^o(\cdot)$, is a function only of the vector difference $\mathbf{s} - \mathbf{u}$. Intrinsically stationary processes are more general than *second-order stationary* processes, for which (2.5) is assumed and assumption (2.6) is replaced with

$$\text{cov}[Z(\mathbf{s}), Z(\mathbf{u})] \equiv C(\mathbf{s}, \mathbf{u}) = C^o(\mathbf{s} - \mathbf{u}), \quad (2.7)$$

where $C^o(\cdot)$ is a function only of the vector difference $\mathbf{s} - \mathbf{u}$.

2.1.3 Ordinary Kriging

Assume that $Z(\cdot)$ is first-order stationary (i.e., assume (2.5)). Suppose that the data consist of observations $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$ of the process at locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset D$. Let $\mathbf{s}_0 \in D$ be some unsampled location and suppose we wish to predict $Z(\mathbf{s}_0)$. Or, more generally, suppose we wish to predict

$$Z(B) \equiv \int_B Z(\mathbf{u}) d\mathbf{u} / |B|; \quad B \subset D, \quad (2.8)$$

where $|B| \equiv \int_B d\mathbf{u}$, the area of B . Note that B may or may not contain sample locations. The spatial best linear unbiased predictor (BLUP), also known as the *ordinary (block) kriging* predictor, of $Z(B)$ is

$$\hat{Z}(B) = \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i), \quad (2.9)$$

where $\lambda_1, \dots, \lambda_n$ are chosen such that

$$E[\hat{Z}(B)] = E[Z(B)] \equiv \mu \quad (2.10)$$

and they minimize

$$E[Z(B) - \sum_{i=1}^n l_i Z(s_i)]^2 \quad (2.11)$$

with respect to l_1, \dots, l_n .

By expressing $\hat{Z}(B)$ as $\lambda_{ok}(B)'Z$, where $\lambda_{ok}(B)' = (\lambda_1, \dots, \lambda_n)$ and $Z = (Z(s_1), \dots, Z(s_n))'$, it is not difficult to show (e.g., Cressie, 1993a, p. 142) that

$$\lambda_{ok}(B)'Z = \left(\gamma(B) + \left\{ \frac{1 - 1'\Gamma^{-1}\gamma(B)}{1'\Gamma^{-1}1} \right\} 1 \right)' \Gamma^{-1}Z, \quad (2.12)$$

where 1 is an $n \times 1$ vector of n ones, Γ is an $n \times n$ matrix whose (i, j) th element is $\gamma(s_i, s_j)$, $\gamma(B) \equiv (\gamma(B, s_1), \dots, \gamma(B, s_n))'$, and $\gamma(B, s_i) \equiv \int_B \gamma(u, s_i) du / |B|$; $i = 1, \dots, n$. The subscript “ok” on $\lambda_{ok}(B)$ emphasizes that we are considering the ordinary kriging vector of coefficients.

The ordinary kriging predictor can also be expressed in terms of the covariance function $C(\cdot, \cdot)$. Assuming (2.2), (2.3), and (2.5), it can be shown (e.g., Cressie, 1993a, p. 143) that

$$\hat{Z}(B) \equiv \lambda_{ok}(B)'Z = \left(c(B) + \left\{ \frac{1 - 1'\Sigma^{-1}c(B)}{1'\Sigma^{-1}1} \right\} 1 \right)' \Sigma^{-1}Z. \quad (2.13)$$

where $\Sigma \equiv \text{var}(Z)$, an $n \times n$ matrix whose (i, j) th element is $C(s_i, s_j)$, $c(B) \equiv (C(B, s_1), \dots, C(B, s_n))'$, and $C(B, s_i) \equiv \int_B C(u, s_i) du / |B|$; $i = 1, \dots, n$. Note that (2.13) can be written as

$$\hat{Z}(B) = \hat{\mu}_{gls} + c(B)' \Sigma^{-1}(Z - \hat{\mu}_{gls}1), \quad (2.14)$$

where $\hat{\mu}_{gls} \equiv 1'\Sigma^{-1}Z / 1'\Sigma^{-1}1$, is the generalized least squares estimator of μ , and is also the best linear unbiased estimator (BLUE) of μ . In the case where μ is known, optimal linear prediction has been discussed *inter alia* by Graybill (1976, pp. 429–439), who shows that the best linear predictor $p_{opt}(Z)$ has the form:

$$p_{opt}(Z) = \mu + c(B)' \Sigma^{-1}(Z - \mu 1). \quad (2.15)$$

Further, if $Z(\cdot)$ is a Gaussian process, then (2.15) is the best (minimum mean-squared-error) predictor, namely $E(Z(B)|\mathbf{Z})$. In geostatistics, (2.15) is also known as the *simple kriging* predictor. If μ is unknown, then $p_{opt}(\mathbf{Z})$ is not a statistic, in which case the spatial BLUP can be obtained by replacing μ in (2.15) with its BLUE, $\hat{\mu}_{gl\beta}$ (Goldberger, 1962).

The *ordinary (point) kriging* predictor of $Z(\mathbf{s}_0)$ is $\hat{Z}(\mathbf{s}_0) = \lambda_{ok}(\mathbf{s}_0)' \mathbf{Z}$, where $\mathbf{s}_0 \in D$ is typically some unsampled location and has the same form as (2.12), but with $\gamma(B)$ replaced by $\gamma(\mathbf{s}_0) \equiv (\gamma(\mathbf{s}_0, \mathbf{s}_1), \dots, \gamma(\mathbf{s}_0, \mathbf{s}_n))'$. Written in terms of the covariance function, it has the same form as (2.13), but with $\mathbf{c}(B)$ replaced by $\mathbf{c}(\mathbf{s}_0) \equiv (C(\mathbf{s}_0, \mathbf{s}_1), \dots, C(\mathbf{s}_0, \mathbf{s}_n))'$.

Define the *ordinary (block) kriging variance*,

$$\sigma_{ok}^2(B) \equiv E[Z(B) - \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)]^2, \quad (2.16)$$

which is the minimized mean squared prediction error. Note once again that the subscript “ok” on $\sigma_{ok}^2(B)$ emphasizes that we are considering the ordinary kriging variance. This can be expressed more explicitly as

$$\sigma_{ok}^2(B) = -\gamma(B, B) + \gamma(B)' \mathbf{T}^{-1} \gamma(B) - (\mathbf{1}' \mathbf{T}^{-1} \gamma(B) - 1)^2 / (\mathbf{1}' \mathbf{T}^{-1} \mathbf{1}), \quad (2.17)$$

where $\gamma(B, B) \equiv \int_B \int_B \gamma(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} / |B|^2$ and the other terms are as defined in (2.12).

The *ordinary (point) kriging variance* is defined as,

$$\sigma_{ok}^2(\mathbf{s}_0) = \gamma(\mathbf{s}_0)' \mathbf{T}^{-1} \gamma(\mathbf{s}_0) - (\mathbf{1}' \mathbf{T}^{-1} \gamma(\mathbf{s}_0) - 1)^2 / (\mathbf{1}' \mathbf{T}^{-1} \mathbf{1}), \quad (2.18)$$

where $\gamma(\mathbf{s}_0) \equiv (\gamma(\mathbf{s}_0, \mathbf{s}_1), \dots, \gamma(\mathbf{s}_0, \mathbf{s}_n))'$.

If we assume that $Z(\cdot)$ is intrinsically stationary, then $\gamma(\mathbf{s}, \mathbf{u}) \equiv \gamma^o(\mathbf{s} - \mathbf{u})$ in equations (2.12), (2.17), and (2.18). If we assume the stricter condition of second-order stationarity, then $C(\mathbf{s}, \mathbf{u}) \equiv C^o(\mathbf{s} - \mathbf{u})$ in equations (2.13) and (2.14).

Note that $\sigma_{ok}^2(B)$ (or $\sigma_{ok}^2(\mathbf{s}_0)$) does not depend on the data $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$, but only on the sample locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, the prediction region B (or prediction location \mathbf{s}_0), the number n of locations sampled, and the semivariogram γ . This property makes kriging very useful for designing spatial sampling plans (e.g., Cressie et al., 1990), since the data play no role in the search for a good sampling plan; all that is required is an accurately modeled variogram of the spatial process.

2.1.4 Constrained Kriging

Suppose we wish to predict $g(Z(B))$, where $g(\cdot)$ is some nonlinear function. Could we not use $g(\lambda_{ok}(B)'Z)$? Cressie (1993b) concludes that $\lambda_{ok}(B)'Z$ is “too smooth”, resulting in an often unacceptable bias for the predictor $g(\lambda_{ok}(B)'Z)$. A predictor of $g(Z(B))$ with better bias properties, called *constrained kriging* (Cressie, 1993b), follows.

Assume only that $Z(\cdot)$ satisfies (2.2), (2.3), and (2.5). Suppose that $g(\cdot)$ is sufficiently smooth to possess at least two continuous derivatives. Then, by the δ -method, we have

$$\begin{aligned} E[g(Z(B))] &\simeq g\{E(Z(B))\} + g''\{E(Z(B))\}\text{var}(Z(B))/2 \\ &= g(\mu) + g''(\mu)\text{var}(Z(B))/2. \end{aligned} \quad (2.19)$$

Let the form of the predictor of $g(Z(B))$ be $g(\alpha'Z)$ satisfying at least the unbiasedness condition on the original scale,

$$E[\alpha'Z] = E[Z(B)] \equiv \mu. \quad (2.20)$$

Using the δ -method, we obtain

$$\begin{aligned} E[g(\alpha'Z)] &\simeq g\{E(\alpha'Z)\} + g''\{E(\alpha'Z)\}\text{var}(\alpha'Z)/2 \\ &= g(\mu) + g''(\mu)\text{var}(\alpha'Z)/2. \end{aligned} \quad (2.21)$$

Thus, as a predictor of $g(Z(B))$,

$$\begin{aligned} \text{Bias}(g(\alpha'Z)) &= E[g(Z(B))] - E[g(\alpha'Z)] \\ &\simeq g''(\mu)\{\text{var}(\alpha'Z) - \text{var}(Z(B))\}/2 \end{aligned} \quad (2.22)$$

and

$$\begin{aligned} \text{MSE}(g(\alpha'Z)) &= E[g(Z(B)) - g(\alpha'Z)]^2 \\ &\simeq (g'(\mu))^2 E\{\alpha'Z - Z(B)\}^2. \end{aligned} \quad (2.23)$$

Note that if $g(\cdot)$ is linear in Z , then $g''(\mu) \equiv 0$, and $\text{Bias}(g(\alpha'Z)) = 0$.

Equation (2.22) indicates that, for nonlinear (and sufficiently smooth) $g(\cdot)$, in order to obtain an (approximately) unbiased predictor of $g(Z(B))$ of the form $g(\alpha'Z)$, where $\alpha = (\alpha_1, \dots, \alpha_n)'$ is chosen to minimize (2.23), we need to minimize $E[\sum_{i=1}^n \alpha_i Z(s_i) - Z(B)]^2$ with respect to $\alpha_1, \dots, \alpha_n$, subject to the unbiasedness constraint (2.20) and the variance constraint

$$\text{var}(\alpha'Z) = C(B, B) \equiv \int_B \int_B C(u, v) du dv / |B|^2 = \text{var}(Z(B)). \quad (2.24)$$

Note that if $Z(\cdot)$ is a Gaussian process, then $g(\alpha'Z)$ is an unbiased predictor of $g(Z(B))$ for any measurable function g , provided that (2.20) and (2.24) are satisfied (Cressie, 1993b).

On most occasions (see below), this constrained minimization can be carried out, yielding the *constrained (block) kriging* predictor $\lambda_{ck}(B)'Z$. Cressie (1993b) shows that

$$\lambda_{ck}(B)' = \frac{1}{m_2}(\mathbf{c}(B) + m_1 \mathbf{1})'\Sigma^{-1}, \quad (2.25)$$

$$m_1 = \frac{m_2 - \mathbf{1}'\Sigma^{-1}\mathbf{c}(B)}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}, \quad (2.26)$$

$$m_2 = \left\{ \frac{(\mathbf{c}(B)'\Sigma^{-1}\mathbf{c}(B))(\mathbf{1}'\Sigma^{-1}\mathbf{1}) - (\mathbf{1}'\Sigma^{-1}\mathbf{c}(B))^2}{(\mathbf{1}'\Sigma^{-1}\mathbf{1})\text{var}(Z(B)) - 1} \right\}^{\frac{1}{2}}, \quad (2.27)$$

and $\mathbf{c}(B)$ and Σ are defined as in (2.13). The numerator of (2.27) is well defined by the Cauchy-Schwarz inequality and the denominator is well defined if

$$\text{var}(Z(B)) > (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1} = \text{var}(\hat{\mu}_{gls}), \quad (2.28)$$

where $\hat{\mu}_{gls}$ is the generalized least squares estimator of μ .

The constrained kriging predictor can also be expressed as

$$\lambda_{ck}(B)'Z = \hat{\mu}_{gls} + \left\{ \frac{C(B, B) - \text{var}(\hat{\mu}_{gls})}{\text{var}(\mathbf{c}(B)'\Sigma^{-1}(Z - \hat{\mu}_{gls}\mathbf{1}))} \right\}^{\frac{1}{2}} \mathbf{c}(B)'\Sigma^{-1}(Z - \hat{\mu}_{gls}\mathbf{1}), \quad (2.29)$$

and it is the best predictor in the class of linear, unbiased, and variance-matching predictors.

The *constrained (block) kriging variance* is

$$\sigma_{ck}^2(B) = 2C(B, B) - 2\lambda_{ck}(B)'\mathbf{c}(B), \quad (2.30)$$

where $C(B, B)$ is given in (2.24).

The *constrained (point) kriging* predictor of $Z(s_0)$ is $\hat{Z}(s_0) = \lambda_{ck}(s_0)'Z$, for some $s_0 \in D$, and it has the same form as (2.25), but with $c(B)$ replaced by $c(s_0)$ and $C(B, B)$ replaced by $C(s_0, s_0)$. Similarly, the constrained (point) kriging variance is

$$\sigma_{ck}^2(s_0) = 2C(s_0, s_0) - 2\lambda_{ck}(s_0)'c(s_0) . \quad (2.31)$$

Note that the constrained kriging predictor is unlikely to perform as well as the kriging predictor if $g(\cdot)$ is linear (e.g., if we wish to predict $g(Z(B)) \equiv Z(B)$), especially if $\text{var}(Z(B))$ and $\text{var}(\lambda_{ok}(B)'Z)$ are substantially different.

2.1.5 Trend

Suppose the data Z are generated by a spatial model with trend, that is, nonconstant $\mu(\cdot)$ in (2.1). When $\mu(\cdot)$ is linear in explanatory variables, the ordinary kriging predictor may be generalized to yield the *universal kriging* predictor (see Cressie, 1993a, Section 3.4). Alternatively, $\mu(\cdot)$ may be estimated nonparametrically (e.g., by median polish; see Cressie, 1993a, Section 3.5), subtracted from the data, and ordinary kriging can then be applied to the residuals. However, the two components $\mu(\cdot)$ and $S(\cdot)$ are not observed individually, so it can happen that a part of $\mu(\cdot)$ is inadvertently included as part of the small-scale stochastic component $S(\cdot)$. In that case, there may be “leakage” of part of the trend into the (estimated) covariance function (2.3) or the (estimated) variogram function (2.4). An example of this is given in Section 3.4 below.

2.1.6 Measurement Error

Data from a spatial process are usually contaminated with measurement error, for which the following model is useful

$$Z(s) = S(s) + \epsilon(s); \quad s \in D, \quad (2.32)$$

where $\epsilon(\cdot)$ represents a zero-mean, white-noise measurement-error process, and

$$S(s) = \mu(s) + \delta(s); \quad s \in D, \quad (2.33)$$

where $\mu(\cdot)$ and $\delta(\cdot)$ are defined as in (2.1); the ϵ and S processes are assumed to be independent. Note that if two observations are taken at a single location, that is, if $Z_1(\mathbf{s})$ and $Z_2(\mathbf{s})$ are observed, they differ from one another only in their error terms, $\epsilon_1(\mathbf{s})$ and $\epsilon_2(\mathbf{s})$, respectively.

The S -process is sometimes referred to as the “state” process or the “signal”, to which measurement errors are added yielding the “noisy” Z -process. It is very important to realize that now we are interested in predicting the “noiseless” S -process over D , but what we actually measure are noisy $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$.

The form of the ordinary kriging and constrained kriging predictors given by (2.13) and (2.25), respectively, do not change under the measurement error model (2.32), except when predicting back at a data location (Cressie, 1993a, p. 128). Further, note that under model (2.32), we have

$$\Sigma \equiv \text{var}(\mathbf{Z}) = \text{var}((S(\mathbf{s}_1), \dots, S(\mathbf{s}_n))') + \tau^2 \mathbf{I}, \quad (2.34)$$

where $\text{var}(\epsilon(\mathbf{s})) \equiv \tau^2$ and \mathbf{I} is the $n \times n$ identity matrix. Thus, equation (2.34) enables the predictors (2.13) and (2.25) to “filter out” the measurement error from the data.

2.1.7 Estimating and Modeling the Variogram

In practice, the variogram 2γ is seldom known and is usually estimated by some nonparametric estimator, such as

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2, \quad (2.35)$$

where $N(\mathbf{h}) \equiv \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\}$ and $|N(\mathbf{h})|$ is the number of distinct ordered pairs in the set $N(\mathbf{h})$. A robust alternative estimator (Cressie, 1993a, p. 75) is

$$2\bar{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \left\{ \sum_{N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}} \right\}^4 / \{0.457 + 0.494/|N(\mathbf{h})|\}. \quad (2.36)$$

Note that these variogram estimators are functions only of the vector difference, $\mathbf{h} = \mathbf{u} - \mathbf{v}$, implicitly assuming that (2.6) holds. If $Z(\cdot)$ is intrinsically stationary, (i.e., both (2.5) and (2.6) hold), then $2\hat{\gamma}$ is an unbiased estimator of $2\gamma^\circ$.

The distinction we made between $\gamma(\cdot, \cdot)$ as a function of two vectors in (2.4), and $\gamma^o(\cdot)$ as a function of vector differences in (2.6) should now be clear: Intrinsic stationarity is required for (unbiased) estimation of the variogram, and not for the kriging equations to be valid. Hereafter we shall be concerned only with variograms and covariances as functions of vector differences, and we shall notate them simply as $2\gamma(\cdot)$ and $C(\cdot)$, respectively.

The nonparametrically estimated variogram cannot be used satisfactorily in (2.12) or (2.13), nor to obtain the kriging variance σ_{ok}^2 , for, among other reasons covered by Cressie (1993a), the estimates are not conditionally negative definite. Moreover, 2γ is estimated only at the lags corresponding to the set of all pairs among the data locations, and these may or may not coincide with the lags required to predict $Z(B)$. Hence, the usual practice is to fit a model $2\gamma(\mathbf{h}; \boldsymbol{\theta})$, whose form is known (apart from a few parameters $\boldsymbol{\theta}$), to $2\hat{\gamma}(\mathbf{h})$ (or $2\bar{\gamma}(\mathbf{h})$). Thus, we use $2\gamma(\cdot; \hat{\boldsymbol{\theta}})$ in place of $2\gamma(\cdot)$ to obtain the kriging predictor and σ_{ok}^2 . (For further discussion on candidate models for $2\gamma(\cdot; \boldsymbol{\theta})$, see Cressie, 1993a, p. 66.) In the case of second-order stationarity, given by (2.5) and (2.7), the stationary covariance function can be obtained from $C(\mathbf{h}) = C(0) - \gamma(\mathbf{h})$.

Assuming the measurement-error model (2.32), the measurement error τ^2 can be estimated from multiple samples at selected sites, if they are available. However, for spatial phenomena, this may not always be possible (e.g., once a soil core has been taken from the ground, it is gone!) But, it may be possible to take extra samples sufficiently close together to avoid contamination by a “microscale” process. If we assume that $\delta(\cdot)$ is L_2 -continuous (i.e., $E[(\delta(\mathbf{s} + \mathbf{h}) - \delta(\mathbf{s}))^2] \rightarrow 0$ as $\|\mathbf{h}\| \rightarrow 0$), then τ^2 can be estimated by the “nugget effect” of the modeled variogram, where the nugget effect \hat{c}_0 is defined as:

$$\hat{c}_0 = \lim_{\|\mathbf{h}\| \rightarrow 0} \gamma(\mathbf{h}, \hat{\boldsymbol{\theta}}). \quad (2.37)$$

2.2 Survey-Sampling Theory

We now present a very brief summary of some of the elements of survey-sampling theory. For more details, the reader is referred to Särndal et al. (1992) and Cochran (1977).

2.2.1 Finite-Population Sampling

Consider a population of N labels which, for convenience, will be represented by the *finite* set $D_f = \{s_1, s_2, \dots, s_N\}$. Associated with each label s_j is a real number $Z(s_j)$, a value of the *study variable* Z corresponding to that label. We assume for the moment that all the elements of the *parameter vector* $\mathbf{Z} = (Z(s_1), \dots, Z(s_N))'$ are fixed and can be obtained without error. The parameter vector \mathbf{Z} may also be referred to as the *target population*.

We are usually interested in making inferences about some numerical summary of \mathbf{Z} , in the form of a *finite population parameter* $\theta(\mathbf{Z})$. Common examples of $\theta(\mathbf{Z})$ include the population size, $N \equiv \sum_{s_j \in D_f} 1$; the population total, $T = \sum_{s_j \in D_f} Z(s_j)$; the population mean, $Z(D_f) \equiv N^{-1}T$; and the population variance, $S_N^2 = (N-1)^{-1} \sum_{s_j \in D_f} (Z(s_j) - Z(D_f))^2$. Other examples include the population cumulative distribution function (CDF), defined as,

$$F(z) \equiv N^{-1} \sum_{s_j \in D_f} I(Z(s_j) \leq z), \quad z \in \mathbb{R}, \quad (2.38)$$

where $I(\cdot)$ is the indicator function (i.e., $I(A) = 1$ if A is true and $I(A) = 0$ if A is not true); and the inverse function of the CDF, the quantile function, defined as

$$q(\alpha) \equiv \inf\{z : F(z) \geq \alpha\}, \quad \alpha \in [0, 1]. \quad (2.39)$$

Suppose that a subset of labels is selected from D_f , randomly or otherwise. This set $A \subset D_f$, is called a *sample*; $s \in A$ is called a *sampling unit*. The process of drawing the sample and obtaining the corresponding Z -values is referred to as a *survey sample*. The *data* collected in a survey sample consist of both the labels and their corresponding measurements, written as

$$X = \{(s_j, Z(s_j)) : s_j \in A\} \quad (2.40)$$

A *sampling design* (or *design*) is a probability mass function $p(\cdot)$ defined on subsets of D_f , such that $\Pr(A = a) = p(a)$. This defines the probability that the sample a is selected. If $p(a) = 1$, for some $a \in D_f$, then the design has no randomization and is said to be *purposive*; if $p(D_f) = 1$, then the design is a *census*.

Define the random indicator function

$$I_j = \begin{cases} 1 & \text{if } s_j \in A \\ 0 & \text{if } s_j \notin A. \end{cases}$$

This is called the *sample membership indicator* of element s_j . The probability that element s_j is included in a sample is given by the *first-order inclusion probability* of element s_j as follows:

$$\pi_j = \Pr(s_j \in A) = \Pr(I_j = 1) = \sum_{a: s_j \in a} p(a). \quad (2.41)$$

The probability that both elements s_i and s_j are included in a sample is given by the *second-order inclusion probability* of s_i and s_j as follows:

$$\pi_{ij} = \Pr(s_i \in A \text{ and } s_j \in A) = \Pr(I_i I_j = 1) = \sum_{a: s_i, s_j \in a} p(a). \quad (2.42)$$

Note that $\pi_{jj} = \pi_j$.

A *probability sampling design* is sometimes defined (e.g., Särndal et al., 1992, p. 32) as a design for which

$$\pi_j > 0, \text{ for all } s_j \in D_f. \quad (2.43)$$

However, Overton (1993) suggests that a probability sampling design should be defined as a design for which (2.43) holds, *and* for which

$$\pi_j \text{ is known, for all } s_j \in A. \quad (2.44)$$

In this chapter, we shall use Overton's (1993) "stronger" definition of a probability sampling design, because it explicitly (and correctly) demands knowledge of the inclusion probabilities for the sample.

A sample a realized by a probability sampling design is called a *probability sample* (or *p-sample*). If for some reason the inclusion probabilities are separated from the sample a , then, by (2.44), a no longer qualifies to be called a *p-sample*. A probability sampling design p has three desirable properties: (i) it eliminates selection bias; (ii) it is objective (as opposed to the purposive selection of "representative" elements or the haphazard selection of convenient elements); and, in particular, (iii) statistical inferences can be made about $\theta(\mathbf{Z})$ based on

the probability structure provided by p , without having to appeal to any statistical model from which \mathbf{Z} is assumed to be a realization. Such inferences are referred to as *design-based* inferences.

If p is a probability sampling design, we can obtain an unbiased estimator of the population total T (see (2.47) below). Further, a *measurable probability sampling design* is sometimes defined (e.g., Särndal et al., 1992, p. 33) as a probability sampling design for which

$$\pi_{ij} > 0, \text{ for all } s_i, s_j \in D_f. \quad (2.45)$$

However, for similar reasons to those given just above, we shall use the “stronger” definition that a measurable probability sampling design requires, in addition to (2.45), that

$$\pi_{ij} \text{ is known, for all } s_i, s_j \in A. \quad (2.46)$$

For such designs, we can obtain from the sample an unbiased estimator of the sampling variance of (2.47) (see (2.49) below).

If p is a probability sampling design, then the *Horvitz-Thompson estimator* (Horvitz and Thompson, 1952), also called the π -estimator, of the total T , is defined as,

$$\hat{T}_{ht} \equiv \sum_{s_j \in A} \frac{Z(s_j)}{\pi_j} = \sum_{s_j \in D_f} \frac{Z(s_j) I_j}{\pi_j}. \quad (2.47)$$

It is an unbiased estimator of T and its *sampling variance* is given by

$$\text{var}(\hat{T}_{ht}) = \sum_{s_i \in D_f} \sum_{s_j \in D_f} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} Z(s_i) Z(s_j). \quad (2.48)$$

If p is a measurable probability sampling design, an unbiased estimator of this variance is

$$\hat{V}_{ht} = \sum_{s_i \in A} \sum_{s_j \in A} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij} \pi_i \pi_j} Z(s_i) Z(s_j). \quad (2.49)$$

The Horvitz-Thompson estimator can be used to estimate several other “totals” of interest, such as the population size, mean, and variance. For example, the Horvitz-Thompson estimator of the population CDF (2.38) is

$$\hat{F}_{ht}(z) = \frac{1}{N} \sum_{s_j \in A} \pi_j^{-1} I(Z(s_j) \leq z); \quad z \in \mathbb{R}. \quad (2.50)$$

2.2.2 Horvitz-Thompson Estimation for a Continuous Population

Cordy (1993) extends the formulation of the Horvitz-Thompson estimator (2.47), its variance (2.48), and its variance estimator (2.49) to the case where the population (of labels) is continuous and the sampling units are points (e.g., in Euclidean space).

Let $D \subset \mathbb{R}^d$; $d \geq 1$, be a continuous population of labels. Assuming a fixed sample size of n , define the *sample space* \mathcal{A} as

$$\mathcal{A} = \{a \equiv (s_1, \dots, s_n) : s_i \in D; i = 1, \dots, n\};$$

that is, the sample $A \equiv (S_1, \dots, S_n) \in \mathcal{A}$ is an ordered n -tuple of random locations, and A has values in D^n .

A (continuous-population) sampling design is a joint probability density function of A with support in D^n , denoted as

$$f(a); s_i \in D, i = 1, \dots, n.$$

The first-order inclusion probabilities are defined as

$$\pi_j^* = \sum_{i=1}^n f_i(s_j); s_j \in D, j = 1, \dots, n,$$

where $f_i(\cdot)$ is the marginal probability density function of S_i , the i th element of A ; $i = 1, \dots, n$.

The second-order inclusion probabilities are defined as

$$\pi_{ij}^* = \sum_{k=1}^n \sum_{l \neq k}^n f_{kl}(s_i, s_j); s_i, s_j \in D, i, j = 1, \dots, n,$$

where $f_{kl}(\cdot, \cdot)$ is the joint marginal probability density function of (S_k, S_l) , the k th and l th elements of A ; $k, l = 1, \dots, n$, $k \neq l$.

Suppose we wish to estimate the (continuous-population) total

$$T^* = \int_D Z(s) d(s).$$

Cordy (1993) proves that if $\pi_j^* > 0$, for all $s_j \in D$, and $\int_D \frac{1}{\pi_j^*} ds_j < \infty$, then an unbiased estimator of T^* is given by

$$\hat{T}_{ht}^* \equiv \sum_{s_j \in A^*} \frac{Z(s_j)}{\pi_j^*}.$$

where A^* is the *set* whose elements comprise the elements of the n -tuple A . Cordy (1993) shows that the *sampling variance* of \hat{T}_{ht}^* is given by

$$\text{var}(\hat{T}_{ht}^*) = \int_D \frac{(Z(s_j))^2}{\pi_j^*} ds_j + \int_D \int_D \frac{(\pi_{ij}^* - \pi_i^* \pi_j^*)}{\pi_i^* \pi_j^*} Z(s_i) Z(s_j) ds_i ds_j.$$

Finally, Cordy (1993) shows that if, in addition, $\pi_{ij}^* > 0$, for all $s_i, s_j \in D$, then an unbiased estimator of $\text{var}(\hat{T}_{ht}^*)$ is given by

$$\hat{V}_{ht}^* = \sum_{s_j \in A^*} \left(\frac{Z(s_j)}{\pi_j^*} \right)^2 + \sum_{s_i \in A^*} \sum_{s_j \in A^*; s_j \neq s_i} \frac{(\pi_{ij}^* - \pi_i^* \pi_j^*)}{\pi_{ij}^* \pi_i^* \pi_j^*} Z(s_i) Z(s_j).$$

In the rest of the chapter, we shall use instead the finite-population formulation, and compare it to a geostatistical approach, adapted to deal with a finite number of units in the domain of interest.

2.2.3 Superpopulation Models in Survey Sampling

Given that the study variable Z can be measured without error, we have so far assumed that \mathbf{Z} consists of N fixed elements and inference is design-based, that is, based on the randomization scheme imposed on the population of labels D_f . Thus, the probability structure that supports design-based inference is an exogenous or externally imposed one.

Suppose now we assume that our target population is a single realization of the random N -vector \mathbf{Z} , and that the joint distribution of \mathbf{Z} can be described by some model ξ , sometimes called the *superpopulation* model. Superpopulation models are used to extend the basis of inference and to formulate estimators with better properties than purely design-based ones.

For example, assume that a sample A has been drawn. Then inference based on the probability structure provided by the superpopulation model ξ conditional on A , is called *model-based* inference. Clearly, model-based inference requires that the model be well specified, that is, that the (model-based) inferences be consistent (in the sense of Fisher consistency; Fisher, 1956, p. 143) with the target population (Overton, 1993).

Superpopulation models are also invoked to formulate estimation methods that may perform substantially better than purely design-based estimation methods if the model is well

specified, and no worse if the model is not well specified (e.g., the regression estimator discussed by Särndal et al., 1992, p. 225). Such methods are said to be *model-assisted*, but not model dependent. For unconditional inference, the probability structures of both p and ξ are used.

2.2.4 Measurement Error

The assumption that the elements of the observable vector $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_N))'$ are free of measurement error may be unrealistic. Suppose, more realistically, that

$$Z(\mathbf{s}_i) = S(\mathbf{s}_i) + \epsilon(\mathbf{s}_i); \quad i = 1, \dots, N, \quad (2.51)$$

where $S(\mathbf{s}_i)$ is the “true” (fixed) value of the i th element of the study population and $\epsilon(\mathbf{s}_i) \equiv Z(\mathbf{s}_i) - S(\mathbf{s}_i)$ represents the i th observational error.

Assuming the measurement error model (2.51), we need to place stochastic structure on the error term in (2.51) if we wish to make any statistical statements about estimators of $\theta(\mathbf{s})$, where $\mathbf{s} \equiv (S(\mathbf{s}_1), \dots, S(\mathbf{s}_N))'$ is the target population.

Let $\hat{\theta}$ be an estimator of $\theta(\mathbf{s})$ and assume some stochastic model for the error process $\epsilon(\cdot)$. Then the estimation error $\hat{\theta} - \theta(\mathbf{s})$ is a random variable whose probability distribution is determined jointly by the sampling design p and by the error process $\epsilon(\cdot)$.

2.2.5 Estimation in the Presence of Measurement Error

For simple error-process models, population-total and population-mean estimation is straightforward. For example, assume model (2.51) and suppose ϵ is a zero-mean, white-noise process with $\text{var}(\epsilon) \equiv \tau^2$. Consider the “true” population total $T_S \equiv \sum_{\mathbf{s}_j \in D_f} S(\mathbf{s}_j)$ and the Horvitz-Thompson estimator $\hat{T}_{ht} = \sum_{\mathbf{s}_j \in A} \pi_j^{-1} Z(\mathbf{s}_j)$. Assume $\epsilon(\cdot)$ and the design p are independent and define the joint expectation

$$E_{p\epsilon}(\cdot) \equiv E_p[E_\epsilon(\cdot|A)],$$

where $E_p(\cdot)$ denotes expectation with respect to p , and $E_\epsilon(\cdot|A)$ is the expectation with respect to ϵ , conditional on the sample A . Then

$$\begin{aligned} E_{p\epsilon}(\hat{T}_{ht}) &= E_p[E_\epsilon(\sum_{s_j \in A} \pi_j^{-1} Z(s_j)|A)] \\ &= E_p[\sum_{s_j \in A} \pi_j^{-1} S(s_j)] = T_S, \end{aligned}$$

and it can be shown (Särndal et al., 1992, ch. 16) that

$$MSE_{p\epsilon}(\hat{T}_{ht}) = V_1 + V_2, \quad (2.52)$$

where $V_1 = \sum_{s_i \in D_f} \sum_{s_j \in D_f} (\pi_{ij} - \pi_i \pi_j) \frac{S(s_i)S(s_j)}{\pi_i \pi_j}$, and $V_2 = \tau^2 \sum_{i=1}^N \pi_i^{-1}$. Thus, the Horvitz-Thompson estimator \hat{T}_{ht} is unbiased for T_S and its variance can be simply partitioned into sampling error and measurement error components.

By contrast, cumulative distribution function (CDF) estimation is not so straightforward, even if the simple error model given above is assumed. Consider the CDF of S ,

$$F_S(z) \equiv \frac{1}{N} \sum_{s_j \in D_f} I(S(s_j) \leq z); \quad z \in \mathbb{R}, \quad (2.53)$$

and consider the Horvitz-Thompson estimator of $F_S(z)$ given by (2.50). It is easy to show that $E_p(\hat{F}_{ht}(z)) = F_Z(z) \equiv \frac{1}{N} \sum_{s_j \in D_f} I(Z(s_j) \leq z)$. Assuming that $\epsilon(\cdot)$ is a zero-mean, white-noise process with $\text{var}(\epsilon) \equiv \tau^2$, we have

$$\begin{aligned} E_{p\epsilon}(\hat{F}_{ht}(z)) &= E_\epsilon(F_Z(z)) = \frac{1}{N} \sum_{s_j \in D_f} G_\epsilon(z - S(s_j)) \\ &\equiv (G_\epsilon * F_S)(z) \neq F_S(z), \text{ if } \tau^2 > 0, \end{aligned}$$

where $G_\epsilon(z) \equiv \Pr(\epsilon(s_j) \leq z)$ and “*” denotes convolution.

Stefanski and Bay (1996) use a simulation-extrapolation deconvolution argument to provide a bias-adjusted CDF estimator. They assume model (2.51) where S is fixed and $(\epsilon(s_1), \dots, \epsilon(s_N)) \sim NI(0, \tau^2)$ (i.e., each $\epsilon(s_i)$ is independently distributed as a $N(0, \tau^2)$ random variable), independent of the sampling design p . Denote the CDF of $\epsilon(s)$ as $\Phi_\tau(\cdot)$. They show that $E_{p\epsilon}[\hat{F}_{ht}(z)] = (\Phi_\tau * F_S)(z)$, which is a special case of the previous result. In order to obtain a deconvoluted CDF estimator, the simulation-extrapolation method of Cook and Stefanski

(1994) was followed, where: (i) additional pseudo random errors of known variance are added to the original data, creating error-inflated “pseudo data” sets; (ii) “pseudo CDF estimators” are recomputed from the pseudo data; and (iii) a trend is established, of the pseudo estimators as a function of the variance of the added errors, and extrapolated backwards to the case of no measurement error.

More specifically, suppose $\eta = (\eta_1, \dots, \eta_m)'$, where the error variance of the i th pseudo random variable is $\tau^2(1 + \eta_i)$, with $\eta_1 < \dots < \eta_m$. Stefanski and Bay (1996) obtain the generic pseudo CDF estimator,

$$\hat{F}_{Z,\eta}(z) = \frac{1}{N} \sum_{i=1}^n \frac{1}{\pi_i} \Phi \left(\frac{z - Z(s_i)}{\tau(1 + \eta)^{\frac{1}{2}}} \right), \quad (2.54)$$

where $\Phi(\cdot)$ is the standard normal CDF. The method they propose depends on the fact that the expectation of (2.54) can be well approximated by a quadratic function in η , that is, by $\beta_0 + \beta_1 \eta + \beta_2 \eta^2$, where $\{\beta_0, \beta_1, \beta_2\}$ are unknown but estimable using linear-model theory. Fix a z in $F_S(z)$. Extrapolating backwards to the case of zero measurement error is equivalent to letting $\eta \rightarrow -1$, resulting in the measurement-error free estimator $\hat{F}_S(z) = \hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2$. From linear-model theory, this can be expressed as

$$\hat{F}_S(z) = \mathbf{a}'(D'D)^{-1}D'\mathbf{v}, \quad (2.55)$$

where $\mathbf{a}' = (1, -1, 1)$, $D = (1, \eta, \eta^2)$, $\mathbf{1} = (1, \dots, 1)'$, $\eta = (\eta_1, \dots, \eta_m)'$, $\eta^2 = (\eta_1^2, \dots, \eta_m^2)$, and $\mathbf{v} = (\hat{F}_{Z,\eta_1}(z), \dots, \hat{F}_{Z,\eta_m}(z))'$ whose elements are given by (2.54).

The authors note that when applying this procedure to the range $z_1 < z_2 < \dots < z_k$, $\hat{F}_S(z)$ may not be monotonic in z , and it is possible that $\hat{F}_S(z) \notin [0, 1]$. As a solution to these problems, they propose fitting an unweighted isotonic regression model to the point estimates $\hat{F}_S(z_i)$; $i = 1, \dots, k$, and truncating the lower and upper ends of the fitted model at 0 and 1, respectively, if necessary.

It is suggested that the η -values be taken equally-spaced over the interval $[0.05, 2]$, and that $m \geq 5$. The authors note that the theory underlying simulation-extrapolation assumes that τ^2 is “small”. Even so, they suggest this method should significantly reduce bias when τ^2 is moderate or even large.

Fuller (1995) invokes a superpopulation model to obtain an estimator of the quantile function (2.39). Using the same superpopulation model, we can derive a CDF estimator of $F_S(z)$ with better bias properties than (2.50). Fuller (1995) assumes the measurement error model (2.51), where $(\epsilon(s_1), \dots, \epsilon(s_N)) \sim NI(0, \tau^2)$, $(S(s_1), \dots, S(s_N)) \sim NI(\mu, \sigma^2)$, and the ϵ and S processes are independent. By invoking this model we get

$$F_S(z) = \Phi\left(\frac{z - \mu}{\sigma}\right), \quad (2.56)$$

where $\Phi(\cdot)$ is the standard normal CDF. Suppose a sample of size n is taken, such that the labels (s_1, \dots, s_k) , with $k < n$, are selected, and n_j replicate samples are taken at label s_j ; $j = 1, \dots, k$, where $\sum_{j=1}^k n_j = n$. Define

$$\hat{\mu} \equiv \frac{1}{k} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} Z_j(s_i).$$

From the analysis of variance of the random-effects model, Fuller (1995) obtains the following variance estimator:

$$\hat{\sigma}^2 \equiv \frac{1}{r} \left(\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{Z}(s_i) - \hat{\mu})^2 - \hat{\tau}^2 \right),$$

where $\bar{Z}(s_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_j(s_i)$, $r = (k-1)^{-1}(n - n^{-1} \sum_{i=1}^k n_i^2)$, and $\hat{\tau}^2 = (\sum_{i=1}^k (n_i - 1))^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (Z_j(s_i) - \bar{Z}(s_i))^2$. It is straightforward to show that $E(\hat{\mu}) = \mu$ and $E(\hat{\sigma}^2) = \sigma^2$. Substituting $\hat{\mu}$ for μ and $\hat{\sigma}$ for σ in (2.56) we obtain the following CDF estimator:

$$\hat{F}_S(z) = \Phi\left(\frac{z - \hat{\mu}}{\hat{\sigma}}\right). \quad (2.57)$$

Nusser et al. (1996) extend this approach to non-Gaussian S and ϵ by assuming that a transformation exists such that both S and ϵ , suitably transformed, are normally distributed.

2.2.6 Inference in Spatial Sampling

Consider the spatial-process model (2.1), and consider the set

$$X = \{(s, Z(s)) : s \in A\} \quad (2.58)$$

where A is a finite (possibly random) subset of $D \in \mathbb{R}^2$. (In practice, D is bounded and is often discretized to a finite grid of locations D_f .) Spatial-model-based inference is based upon the probability structure ξ defined by (2.1), conditional on the sample A of locations. If A is selected by a probability sampling design p , then spatial design-based inference is supported by the probability structure defined by p , conditional on $Z(\cdot)$.

Now assume the spatial model (2.32) that includes measurement error. Spatial-model-based inference is similar to that for model (2.1), but now ξ also includes the measurement-error component $\epsilon(\cdot)$. Further, for (2.32), spatial design-based inference is supported by the design p and the error process $\epsilon(\cdot)$ (see Sections 2.2.4 and 2.2.5), conditional on $S(\cdot)$. Here, both model-based and design-based inferences depend on the probability structure defined by $\epsilon(\cdot)$.

To avoid confusion, we wish to clarify some terminology: In the spatial-model context, interest is usually in the “prediction” of quantities assumed to be random (i.e., functions of $S(\cdot)$); in the survey-sampling context, those same quantities (i.e., functions of $(S(s_1), \dots, S(s_N))'$), may be assumed fixed and consequently the corresponding inference is termed “estimation”. In our work, we shall generally use the term “prediction” for making inference on $S(\cdot)$ or $(S(s_1), \dots, S(s_N))'$, regardless of whether it is model-based or design-based.

2.3 Review of the Spatial Sampling Literature

The past literature on spatial sampling has been concerned with a number of issues related to choosing $A \equiv \{s_1, \dots, s_n\}$ and $n = |A|$ in a spatial domain of interest. The first issue is that of model-based versus design-based sampling approaches. The second issue deals with design criteria by which the performance of a spatial design is assessed, for a given predictor. The third is one of comparing various popular, though not necessarily optimal, designs, and the fourth relates to adaptive sampling procedures.

2.3.1 Model-Based and Design-Based Sampling Approaches

Model-based and design-based approaches to spatial-sampling theory each depend on a different source of probability structure upon which inferences may be made (Section 2.2.3

and 2.2.6). What emerges from the literature is that design-based inference may be more robust than model-based inference, but that an appropriate model-based analysis may perform substantially better, provided that the model-based inferences are (Fisher) consistent with the target population (Overton, 1993).

The central problem with model-based inference is that if the model is not consistent with the target population, then a purely model-based analysis may yield substantially biased estimates of population parameters and very misleading estimates of sampling precision (McArthur, 1987; Overton, 1993). This suggests that the design-based approach may be more appropriate if a consistent spatial model cannot be identified reliably. De Gruijter and ter Braak (1990) argue that, although design-based efficiency may be less than the optimal model-based efficiency, such a loss may be a worthwhile premium to pay for robustness against model errors and for achieving p -unbiasedness (i.e., design-unbiasedness).

Some model-based methods incorporate design-based approaches (Cox et al., 1997). Overton (1993) says that model-based inference can improve precision of a p -sample, often greatly, and is likely to be consistent with the target population. Cressie and Ver Hoef (1993) and Stevens (1994) suggest that model-based inference may be strengthened by characteristics of a p -sample (e.g., a systematic design with random starting point), by eliminating selection bias in the choice of the sample locations, and by providing a mechanism for inferences free of the assumed spatial model. However, Cressie and Ver Hoef (1993) also suggest that purely design-based inferences, ignoring models describing small-scale spatial dependence structures, are severely limited in the range of questions they can address. Thus, armed with flexible spatial models and good model diagnostics, the use of spatial models can greatly enhance the science of spatial sampling.

2.3.2 Design Criteria

A central problem in sampling theory is the search for an optimal sampling *strategy*, that is, the search for a design/predictor combination that best achieves our objectives. Prediction objectives commonly include point and block prediction (see Section 2.1) and Cox et al. (1997)

list several other prediction objectives in the spatial context. These include: the prediction of the average of a nonlinear function of the spatial process $S(\cdot)$; the prediction of the maximum of $S(\cdot)$; and the prediction of the subregion for which the spatial process exceeds a given threshold.

In much of the spatial sampling literature, the major concern seems to be about design optimization for a specified predictor. That is, for a given predictor, *design* criteria are usually considered. This discussion will concentrate largely on model-based design criteria, mainly because of their particular applicability to the spatial sampling situation. Design-based design criteria (e.g., Horvitz-Thompson sampling variances or MSEs with respect to the sampling design) are, as Cox et al. (1997) note, elementary from a statistical point of view and can be obtained from any reasonable text on survey sampling (e.g., Särndal et al. (1992) and Cochran (1977)).

So, given some measurement-error spatial-process model and the accompanying spatial analyses discussed in Section 2.1, various model-based design criteria can be formulated to assess the performance of different sampling designs for given predictors. Considering only linear unbiased predictors $\{\hat{S}(\mathbf{u}; \mathbf{Z}) : \mathbf{u} \in A\}$, which are functions of $X = \{(s_1, Z(s_1)), \dots, (s_n, Z(s_n))\}$, Cox et al. (1997) enumerate three sampling-design criteria:

(D-1) The Integrated Mean Squared Error (IMSE) criterion: Minimize with respect to sampling locations $\{s_1, \dots, s_n\}$,

$$IMSE(s_1, \dots, s_n) \equiv \int_D E[\hat{S}(\mathbf{u}; \mathbf{Z}) - S(\mathbf{u})]^2 d\mathbf{u} .$$

(D-2) The Maximum MSE (MMSE) criterion: Minimize with respect to $\{s_1, \dots, s_n\}$,

$$MMSE(s_1, \dots, s_n) \equiv \sup_{\mathbf{u} \in D} E[\hat{S}(\mathbf{u}; \mathbf{Z}) - S(\mathbf{u})]^2$$

(D-3) The Entropy Criterion: Maximize with respect to $\{s_1, \dots, s_n\}$,

$$H(s_1, \dots, s_n) \equiv E[-\log \{f(Z(s_1), \dots, Z(s_n))\}] ,$$

where f denotes the probability density function of $(Z(s_1), \dots, Z(s_n))$.

Note that, for ordinary kriging (constrained kriging) predictors, criteria (D-1) and (D-2) simplify to $\int_D \sigma_{ok(ck)}^2(u) du$ and $\sup_{u \in D} \{\sigma_{ok(ck)}^2(u)\}$, respectively.

Cressie (1993a) modifies (D-1) and (D-2) by including weight functions in the integrals. He gives as an example of weights, indicator functions that focus attention on subregions whose mean and variance exceed some given threshold.

Clearly, it makes sense to choose criteria that relate to the major objectives in the study. Thus, (D-1) is an appropriate criterion if a design is needed to perform “best on average”, and (D-2) is useful if one seeks to minimize the worst case. The entropy criterion (D-3) is claimed by Caselton and Zidek (1984) and Guttorp et al. (1993) to be useful in studies with multiple objectives. It is usually considered only in cases where finitely many potential locations are available, but this need not be a limitation, since a very fine grid of possible locations is still finite in number. A feature of (D-3) is that if $Z(\cdot)$ is a Gaussian process, the maximization of H is equivalent to the maximization of the determinant of the covariance matrix of $(Z(s_1), \dots, Z(s_n))$, and hence it is also known as the D-optimality criterion (Cox et al., 1997).

It has been well documented (Olea, 1984 and Cressie et al., 1990), that regular, particularly triangular networks, do well in terms of minimizing maximum σ_{ok}^2 . However, Haas (1992) points out that, in practice, the variogram (2.6) typically has to be estimated and then modeled (as described in Section 2.1.7), so that some clustering of the design points is necessary for good variogram estimation at short lags, where accuracy is usually most important.

Haas (1992) addresses the problem of redesigning a continental-scale monitoring network by providing a method for optimally adding sites to a subregion of the continent, using a bivariate criterion. He defines the relative error estimate (REE) of a region B as $\hat{\sigma}_{ok}(B)/\hat{Z}(B)$, where the (estimated) kriging standard error is obtained from (2.17), but with $\gamma(\cdot)$ replaced by $\gamma(\cdot; \hat{\theta})$. The criterion he proposes is to minimize both the REE over the subregion in question and the standard deviation of the REE at the subregion’s center; the latter, he suggests, could be estimated by the sample standard deviation of a simulated sampling distribution of the REE at the subregion’s center.

Guttorp et al. (1993) make use of the entropy criterion (D-3), which they propose as a generic objective designed to meet some quality requirement for all objectives, even though it may not be best for any given objective. Caselton and Zidek (1984) explore the usefulness of the entropy criterion for long-term studies where all possible uses of data are unlikely to be unique or foreseen. They view network optimization as a problem of selecting a set M of monitored site locations so that the increase in information about a set U of unmonitored sites, after observing $\{Z(s) : s \in M\}$, is maximized. Here $D_f = M \cup U$. Thus, if $\mathbf{Z}_M, \mathbf{Z}_U$ are random variable vectors of the process at the monitored and unmonitored sites, respectively, let $f(\mathbf{Z}_M, \mathbf{Z}_U)$ be the joint distribution of the vectors and $f(\mathbf{Z}_M), f(\mathbf{Z}_U)$ be the marginal distributions. If $f(\mathbf{Z}_U)$ is interpreted as a prior density function, then Caselton and Zidek (1984) propose that the locations M should be chosen to maximize what Pérez-Abreu and Rodríguez (1996) refer to as the *Shannon Information Index*,

$$I(U, M) = \iint \log \left\{ \frac{f(\mathbf{Z}_M, \mathbf{Z}_U)}{f(\mathbf{Z}_M)f(\mathbf{Z}_U)} \right\} f(\mathbf{Z}_M, \mathbf{Z}_U) d\mathbf{Z}_U d\mathbf{Z}_M,$$

which is the information about \mathbf{Z}_U contained in \mathbf{Z}_M . Note that if $f(\mathbf{Z}_M)$ and $f(\mathbf{Z}_U)$ are independent, then $I(U, M) = 0$ (i.e., the monitored sites provide no information about the unmonitored sites). Pérez-Abreu and Rodríguez (1996) extend this to the multivariate $\mathbf{Z}(\cdot)$ case.

Haas (1992) observes some problems with this approach: (i) the prior distribution $f(\mathbf{Z}_U)$ may be difficult to specify in practice; (ii) a parametric multivariate distribution needs to be specified for all of the sites, monitored and unmonitored; (iii) the mean and covariance structure of the multivariate distribution must be estimated in practice, thereby introducing error into the computation minimizing $I(U, M)$, and this has not been taken into account in the current version of the theory.

So far, we have been concerned mainly with the precision aspects of model-based design criteria, whether or not the invoked models relate meaningfully to the target population. If a superpopulation model is assumed, we must be clear whether our target population is the “real-world” population (i.e., a single realization of the invoked model), or if it is the superpopulation itself. For example, given the spatial model (2.1), are we interested in predicting

the real-world mean $Z(B)$, or is our interest in estimating the model mean $E(Z(B)) \equiv \mu(B)$? Superpopulation-model parameter estimation may be appropriate if we believe that the model describes some causal or mechanistic behavior, but in the spatial context, interest is more likely to be in the real-world population.

Overton (1993) suggests strongly that if our interest is in the real-world population, then our primary statistical criterion should be that (model-based) inferences be (Fisher) consistent with the real-world population. Sampling designs, which provide precise predictors according to the models but which are not consistent with the real-world population, are simply unacceptable. He further suggests that probability sampling designs best ensure real-world-population consistency.

Certain criteria may be useful either for selecting a good design or for selecting a good predictor, depending on how the optimization is done. For example, suppose inferences about some real-world population are desired, ξ is some assumed superpopulation model and p is a probability sampling design. Then Särndal et al. (1992, p. 516) suggest that the predictor that minimizes the unconditional MSE be selected. For example, suppose we wish to predict $S(B)$ with $\hat{S}(B)$. The criterion to minimize, with respect to choice of predictor $\hat{S}(B)$, is

$$E_{p\xi}[(\hat{S}(B) - S(B))^2], \quad (2.59)$$

where $E_{p\xi}(\cdot)$ is the expectation taken with respect to both p and ξ . If $E_{p\xi}[\hat{S}(B) - S(B)] = 0$, then (2.59) is called the *anticipated variance* (Isaki and Fuller, 1982), and this is commonly used as a design criterion; that is, it is minimized with respect to p , for a given predictor and a given model ξ (e.g., Breidt, 1995a).

In cases where design-based inference is to be emphasized, Overton and Stehman (1993) state that, while (design-based) precision should be a primary design criterion, an important secondary criterion is the ability to obtain adequate variance estimators. Unbiased variance estimators of the form (2.49) exist for Horvitz-Thompson estimators if p is a measurable probability sampling design, but (2.49) may behave badly under certain conditions (e.g., (2.49) may be negative; Särndal et al., 1992, p. 47). However, for those designs that have some pairwise inclusion probabilities that are zero (e.g., systematic sampling designs), adequate

variance estimators are not so obviously obtainable, and models may have to be invoked to derive them.

The sampling optimality criteria presented so far, have all been based on statistical considerations. Frequently, cost or economic considerations provide a very important limitation, which should be introduced into the objective function.

Bras and Rodriguez-Iturbe (1976) propose the objective function,

$$\delta(n, A) + \beta\kappa(n, A), \quad (2.60)$$

where $\delta(n, A)$ is a measure of statistical accuracy, $\kappa(n, A)$ is the cost of sampling, β is a measure of accuracy obtained from a unit increase in cost, n is the sample size, and $A \equiv \{s_1, \dots, s_n\}$ is the set of sample locations. Bras and Rodriguez-Iturbe (1976) suggest how (2.60) can be optimized numerically over sampling locations A and sample size n . Bogardi et al. (1985) consider the problem of optimal spatial sampling design as one of multicriterion decision making. A composite objective function, measuring statistical and economic tradeoffs, is proposed. The optimal *rectangular* network is achieved through compromise programming (Zeleny, 1982). Englund and Heravi (1992) propose the use of conditional simulation (i.e., simulation from a spatial model whose realizations always go through the observations at the sampling locations $\{s_1, \dots, s_n\}$) as a powerful tool for the optimization of economic objective functions.

2.3.3 Comparison of Sampling Designs

The computational problem of obtaining optimal IMSE designs is difficult (Cressie, 1993a, pp. 319, 320 and Cox et al., 1997). But, it is often relatively easy to obtain “good” designs, and standard iterative algorithms usually make substantial improvements over the initial design after only a few iterations (Cox et al., 1997). Cox et al. (1997) further state that, in their experience, good designs tend to spread points uniformly in the design region, echoing the results of Dalenius et al. (1960), Olea (1984), and Currin et al. (1991).

Using model-based design criteria, Olea (1984) and Cressie et al. (1990) compare various popular, but not necessarily optimal designs. Olea (1984) shows that the geometrical con-

figuration of a network is the most important consideration in optimal network design and, in particular, regular triangular grids minimize the maximum σ_{ok}^2 over the spatial domain D . Cressie et al. (1990) illustrate that, for spatial processes with increasing spatial dependence (up to a point), systematic sampling designs (SYS) are better than stratified random sampling designs (STS), which are in turn better than simple random sampling designs (SRS), with respect to a risk function that depends on σ_{ok}^2 . Olea (1984) reaches conclusions that agree with this ranking, adding that clustered designs are by far the worst.

Overton and Stehman (1993) compare designs with respect to two criteria, viz., the relative precision of the designs with respect to design-based criteria as the primary consideration, and the capability for adequate variance estimation of the designs as a second important consideration. They used three types of surface $Z(\cdot)$: planar, quadratic, and sinusoidal. The designs considered were SYS, SRS and another design called tessellation-stratified sampling (TSS), where the hexagons of a triangular-grid tessellation were used as strata and one random point per stratum was selected. They note that TSS has the advantages of a systematic design (evenly distributed samples over the domain of interest) without its disadvantage (inability to handle certain periodicities in the surface). In almost all cases, TSS outperformed SYS with respect to both criteria under consideration, and in some cases greatly so. The sampling design SRS performed worst in all cases.

A study by McArthur (1987) involves sampling over a peaked surface corresponding to environmental contaminants emerging from a point source, where the pollutants are more concentrated near the source. In this case, a preferential grid (stratification of two systematic grids, with one of higher sampling density centered on the location of the peak) is the most accurate and precise by far, but estimation of the design-based sampling variance from one realization of the design is not possible. It was found that STS and importance sampling (IS), a Monte Carlo method for computing integrals (see, e.g., Hammersley and Handscomb, 1964), were very useful for predicting $Z(B)$ if good prior knowledge of the surface $Z(\cdot)$ were available. If not, SYS or SRS is suggested, with preference given to the former if no periodicities are likely to occur in the surface. Estimation of the design-based sampling variance from one realization

of a SYS design is also not possible, but Wolter (1984) suggests approximate methods to get around this.

Markov chain (MC) designs for one-per-stratum sampling are presented by Breidt (1995a) for finite populations, and by Breidt (1995b) for continuous populations. For simplicity, consider for the moment a one-dimensional sampling domain partitioned into equally sized strata. After initially selecting a sampling location from an “end” stratum by means of some probability function, the location-selection procedure moves sequentially along the strata of the domain to the other end, by means of a stochastic process satisfying the Markov property (i.e., the probability of a future selection depends only on the present selection and not on past selections). Thus, the probability of a location being selected depends only on the selection in the immediately preceding stratum. This procedure is easily extended to two dimensions by considering two independent stochastic processes, with one operating on the “latitude” coordinate, and the other on the “longitude” co-ordinate. Given the stochastic processes specified by Breidt (1995b) in the continuous population case, what results is a range of designs that include as special cases SYS, STS with one sampling unit per stratum, and balanced systematic sampling designs (BAS). A balanced design is one in which the sampling location within one stratum and the sampling location within an adjacent stratum are equidistant from the stratum boundary that separates them. Breidt (1995a and 1995b) evaluates MC designs under a variety of superpopulation models, using the *anticipated variance* criterion (see (2.59)). He demonstrates that new designs from within the broader MC class are competitive with the standard designs, SYS, STS, and BAS, under a variety of models. In particular, he shows that for models with a dominant trend term, the optimal design is close to BAS; but if an autocorrelation (spatial dependence) term dominates, then the optimal design shifts closer to SYS.

It is well known that systematic designs are potentially disastrous in the face of periodicities occurring within the sampling domain. It has been noted that while few periodicities occur in nature, human effects on the landscape are often systematic. Much of the Midwest of the USA has imposed upon it a one-mile square grid of gravel roads and, as Breidt (1995a) remarks,

a survey-sampler using a systematic design with a one-mile interval and an unlucky random start, might conclude that Iowa is covered by gravel roads! Such unfortunate occurrences may be avoided by selecting a non-systematic MC design.

2.3.4 Adaptive Sampling

For spatial phenomena that are rare, or clustered, or both, adaptive sampling methods (Thompson, 1992) may be far more useful than traditional sampling methods. Suppose that some variable of interest Z is mostly zero but that its nonzero values are spatially clustered. We wish to (i) estimate the average $Z(D_f)$ or total T of Z ; and (ii) locate the “pockets” where Z is nonzero. Traditional sampling methods provide unbiased estimates of $Z(D_f)$ and T , but with high variance; and the maps they provide detailing the occurrence of pockets of Z are usually highly inaccurate (Seber and Thompson, 1994). The adaptive-sampling procedure goes as follows: (i) select an initial sample by some conventional design (e.g., SRS); (ii) add to the initial sample any units that satisfy some specified condition (e.g., add neighboring units to all initial sample units where Z occurs). Thompson (1992) provides a theory to modify selection probabilities and to obtain consistent estimators, which he shows can be much more efficient than SRS estimators. Cox et al. (1997) suggest that this may be useful for “hot-spot” identification in environmental problems.

3 Computer Simulation Experiment

Consider a spatial domain over which a Gaussian spatial process (superpopulation) model is defined. The aim of this study is to use design-based criteria to compare different analyses (i.e., prediction methods) and different sampling designs under a variety of conditions. In much of the spatial sampling literature, the emphasis is on gauging sampling designs for given predictors; in this study the emphasis is on the choice of prediction method under different sampling designs.

A computer experiment was devised, complete with “factors” and “responses”. The factors of the experiment include the sampling designs, the prediction methods, and the different con-

ditions under which the designs and analyses were conducted. Performance criteria constitute the responses. The details of the spatial-process model and of the computer experiment are now presented.

3.1 Gaussian Spatial Models

The spatial domain D considered in this experiment is a square region in \mathbb{R}^2 which is discretized into $D_f = \{(x, y) : x = 1, \dots, 10; y = 1, \dots, 10\}$, a 10×10 grid of 100 locations. Over this domain, the following measurement-error spatial model was invoked:

$$Z(s) \equiv S(s) + \epsilon(s); \quad s = (x, y) \in D_f, \quad (3.1)$$

where $\epsilon(\cdot)$ is a zero-mean, white-noise measurement-error process, and we define the state process,

$$S(s) = \mu(s) + \delta(s); \quad s = (x, y) \in D_f, \quad (3.2)$$

where $\mu(x, y) = \beta(x - 4.5)$; $x = 1, \dots, 10$; $y = 1, \dots, 10$, with β to be specified in Section 3.2.3. The set of 100 δ -values $\{\delta(x, y) : x = 1, \dots, 10; y = 1, \dots, 10\}$ corresponding to the 100 grid locations $\{(x, y) : x = 1, \dots, 10; y = 1, \dots, 10\}$ were generated according to a zero-mean, Gaussian spatial process model with covariance function

$$\text{cov}(\delta(i, j), \delta(k, l)) = \sigma^2 \begin{cases} \rho^h & \text{for } \rho \in \{0.1, 0.2, \dots, 0.9\} \\ 1 & \text{if } \rho = 0 \text{ and } (i, j) = (k, l) \\ 0 & \text{if } \rho = 0 \text{ and } (i, j) \neq (k, l), \end{cases} \quad (3.3)$$

where $h \equiv \{(i - k)^2 + (j - l)^2\}^{\frac{1}{2}}$ and, without loss of generality, $\sigma^2 = 1$ was chosen.

Note that $\delta(\cdot)$ is a second-order stationary process and, for $\rho = 0$, $(\delta(1, 1), \dots, \delta(10, 10)) \sim NI(0, 1)$. As $\rho \rightarrow 1$ (although $\rho > 0.9$ is out of the range of values used in this study), the δ -process tends towards a common value, that value being distributed as $N(0, 1)$.

The Z -process in (3.1) was obtained by defining the measurement-error process through $(\epsilon(1, 1), \dots, \epsilon(10, 10)) \sim NI(0, \tau^2)$, where τ^2 is specified in Section 3.2.4 below.

3.2 Factors of the Computer Experiment

3.2.1 Regions

Let the entire domain of the 10×10 locations be called the *global* region G . Demarcate a 3×3 subregion of 9 locations $\{(x, y) : x = 1, 2, 3; y = 1, 2, 3\}$ of G , and call this the *local* region L . These are the two regions over which statistical inference of various forms will be made.

3.2.2 Strength of Spatial Dependence

The strength of spatial dependence in the δ -process, characterized by (3.3), ranges from zero to very strong and is indexed by $\rho \in \{0, 0.1, \dots, 0.9\}$.

3.2.3 Trend

Two values were given to the trend term in (3.2), viz., $\beta \in \{0, (8.25)^{-1/2}\}$. Note that for $\beta = 0$, $S(\cdot)$ and $Z(\cdot)$ are both (second-order) stationary Gaussian processes. The choice of the other value of β is discussed below.

3.2.4 Noise

Two levels of measurement error were considered: $\tau^2 = 0.1$ and 2. These two levels will be referred to as *low* noise and *high* noise, respectively. Their choice is discussed below.

The trend and noise parameters were chosen according to a square-root signal-to-noise ratio scale. Define the “signal” variance, σ_s^2 , as the sum of the model variance of the δ -process and the sample variance of the trend process; that is, $\sigma_s^2 = \text{var}_\xi(\delta(s)) + \frac{1}{10} \sum_{x=1}^{10} \beta^2 (x - 4.5)^2 = 1 + 8.25\beta^2$. Thus, $\beta \in \{0, (8.25)^{-1/2}\}$ yields $\sigma_s^2 \in \{1, 2\}$. Combining both levels of β and both levels of τ^2 we obtain the following four ratios: $\sigma_s/\tau \in \{0.71, 1, 3.16, 4.47\}$.

These ratios do not take into account the spatial “ ρ -effect”, (i.e., strength of spatial correlation). As ρ increases, the δ “realization” becomes much smoother, even though $\text{var}_\xi(\delta(s)) \equiv 1$. For large ρ , $\text{var}_\xi(\delta(s))$ accurately represents the variability due to $\delta(\cdot)$ over a very large spatial domain, but over small spatial domains (such as 10×10), where $\delta(\cdot)$ is likely to be much

smoother, the signal variance may appear to be substantially less than 1. Thus the apparent ratio from any given realization may be much smaller than the prespecified σ_s/τ .

3.2.5 Realizations of the State Process

The S -process in (3.2) is held fixed over the randomization component in the design. A vector \mathbf{Y} of 100 $NI(0, 1)$ values were generated, and the 100 values of $\delta(\cdot)$ (i.e., δ -realizations) were obtained from $V(\rho)^{\frac{1}{2}}\mathbf{Y}$ where, from (3.3), the matrix $V(\rho)$ with elements $\text{cov}(\delta(u), \delta(s))$; $u, s \in G$, is a function of $\rho \in \{0.0, 0.1, \dots, 0.9\}$. Thus, a set of *ten* δ -realizations were generated corresponding to the set of *ten* ρ values, using the *same* \mathbf{Y} vector. And, two S -realizations, corresponding to the two values of the trend parameter β , were created from each δ -realization, by adding the appropriate trend term given in (3.2). See Figure 3.1, where S -realizations are shown for the two β -values in question, and for $\rho \in \{0, 0.5, 0.9\}$.

Three different \mathbf{Y} vectors of 100 $NI(0, 1)$ values were used to generate three realizations of the S -process (i.e., S -realizations) for each value of ρ and β .

3.2.6 Sampling Designs

A subset of 20 locations $\{s_1, \dots, s_{20}\}$ was selected according to various sampling designs from among the 100 grid locations, and the corresponding Z -values were generated by adding 20 ϵ -values, independently generated from a $N(0, \tau^2)$ distribution, to the corresponding values from the state process, $\{S(s_j) : j = 1, \dots, 20\}$. A different random sample of locations yields a different set of S -values and a different set of measurement errors but the underlying 100 values $\{S(x, y) : x = 1, \dots, 10; y = 1, \dots, 10\}$ of the S -realization remain the same over randomization of the sampling locations. The sampling designs considered are given as follows:

1. *Systematic Random Sampling* (SYS, notated “Y” on the figures): Consider the design configuration where the s_j -th column has sampling location at $(1, j) + (2j - 2, 0)$ and $(6, j) + (2j - 2, 0)$; $j = 1, \dots, 10$, and componentwise addition is modulo 10. Randomly “start” the 20 samples by choosing the first location to be $(1, 1) + (0, k)$, where k is uniformly distributed on $\{0, 1, 2, 3, 4\}$.

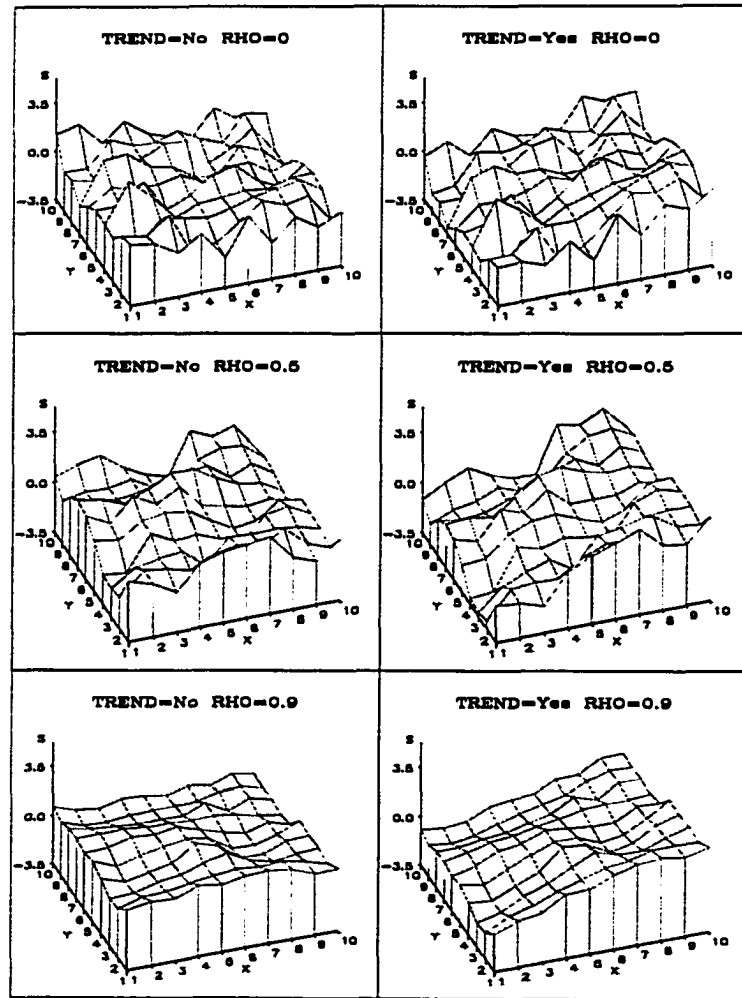


Figure 3.1 Generated S -realizations over global region $G = \{(x, y) : x = 1, \dots, 10; y = 1, \dots, 10\}$, for trend parameter $\beta \in \{0, (8.25)^{-1/2}\}$ and spatial-correlation parameter $\rho \in \{0, 0.5, 0.9\}$.

2. *Stratified Random Sampling* (STS, notated “T” on the figures): 2 locations are chosen randomly (without replacement) from each “column” of the grid.
3. *Simple Random Sampling* (SRS, notated “R” on the figures): 20 locations are chosen randomly (without replacement) from the entire grid.
4. *Clustered* (CLU, notated “C” on the figures): 10 observations are chosen from locations (2,2) and (9,9) each.

The three designs, SYS, STS, and SRS, are probability sampling designs, and CLU is a purposive design. The S -process is stationary within the strata defined by STS, irrespective of the value of the trend parameter β . Because of its desirable properties with respect to spatial model-based criteria (see Section 2.3.2), SYS may be regarded as a “spatial” design.

The sample size was fixed at $n = 20$ and thus is not a factor in the experiment. However, for some prediction methods over the local region (see Section 3.2.7 below), the sample size is a random variable. Note that it is possible that the 20 sampled locations may all fall outside L for STS and SRS designs.

3.2.7 Prediction Methods

Prediction presupposes some target: In this study, interest is in two predictands, the spatial mean and the spatial cumulative distribution function (SCDF), to be defined in what follows.

The Spatial Mean

The spatial mean of $S(\cdot)$ over a region B is defined:

$$S(B) \equiv \sum_{u \in B} S(u) / |B|, \quad (3.4)$$

where $|B|$ is the number of S -values whose locations fall in B . Thus,

$$S(G) = \frac{1}{100} \sum_{x=1}^{10} \sum_{y=1}^{10} S(x, y) \quad (3.5)$$

$$S(L) = \frac{1}{9} \sum_{x=1}^3 \sum_{y=1}^3 S(x, y). \quad (3.6)$$

We will consider four predictors of these two quantities, all of which are functions of $\mathbf{Z} \equiv (Z(s_1), \dots, Z(s_{20}))'$, the data vector corresponding to the 20 sampled locations.

1. *Ordinary Kriging* (OK): Define

$$\hat{S}_{ok}(B) \equiv \lambda_{ok}(B)' \mathbf{Z}; \quad B = G \text{ or } L, \quad (3.7)$$

where $\lambda_{ok}(B)' \mathbf{Z}$ is given by equation (2.13) with $n = 20$. Recall that $\hat{S}_{ok}(B)$ is the *best linear unbiased predictor* (BLUP) of $S(B)$.

2. *Constrained Kriging* (CK): Define

$$\hat{S}_{ck}(B) \equiv \begin{cases} \lambda_{ck}(B)' \mathbf{Z}, & \text{if } \rho > 0, B = G \text{ or } L \\ \bar{Z} + \left\{ \frac{n - |B|(1 + \tau^2)}{|B|(1 + \tau^2) \{ n \sum_{i=1}^m |B_i|^2 - N_m^2 \}} \right\}^{\frac{1}{2}} \\ \quad \times (\sum_{i=1}^m |B_i| Z(\mathbf{v}_i) - N_m \bar{Z}), & \text{if } \rho = 0, B = L \\ \bar{Z}, & \text{if } \rho = 0, B = G \end{cases} \quad (3.8)$$

where $\lambda_{ck}(B)'$ is given by equation (2.25) with $n = 20$. The solution for the case where $\rho = 0$, $B = L$, and an explanation of the terms in (3.8) is given in the Appendix. No solution to the constrained kriging equations exists for $\rho = 0$, $B = G$, so \bar{Z} was used instead.

3. *Regional Poststratification* (RP): Define

$$\hat{S}_{rp}(G) \equiv \begin{cases} \frac{|L|}{|G|}(\bar{Z}_L) + \frac{|G| - |L|}{|G|}(\bar{Z}_{G-L}), & \text{if } n_L > 0 \\ \bar{Z}, & \text{if } n_L = 0, \end{cases} \quad (3.9)$$

where \bar{Z}_L , \bar{Z}_{G-L} , and \bar{Z} are averages of the Z -values contained in L , $G - L$ and G respectively; and n_L is the number of sample locations in L . The local-region predictor is defined

$$\hat{S}_{rp}(L) \equiv \begin{cases} \bar{Z}_L, & \text{if } n_L > 0 \\ \bar{Z}, & \text{if } n_L = 0. \end{cases} \quad (3.10)$$

Note that the sample size specified in (3.10) is a random variable.

4. *Arithmetic Mean* (AM): Define

$$\hat{S}_{am}(B) \equiv \bar{Z}; \quad B = G \text{ or } L. \quad (3.11)$$

It should be noted that $\hat{S}_{am}(G)$ is a Horvitz-Thompson estimator for $S(G)$ for designs with equal first-order inclusion probabilities, that is, for SYS, STS, and SRS. (All pairwise inclusion probabilities for the designs STS and SRS are positive, hence \hat{V}_{ht} given by (2.49) can be obtained for those designs. This is not the case for SYS.)

Observe that the first two predictors, ordinary kriging and constrained kriging, are model-based “spatial” predictors and the last two predictors, regional poststratification and arithmetic mean, are design-based “non-spatial” predictors. The model assumptions of ordinary and constrained kriging are that the spatial covariance parameters are known.

The Spatial Cumulative Distribution Function

Let F_B denote the spatial cumulative distribution function (SCDF) of $\{S(s) : s \in B\}$; specifically,

$$F_B(z) \equiv \frac{1}{|B|} \sum_{u \in B} I(S(u) \leq z); \quad z \in \mathbb{R}, \quad (3.12)$$

where $|B| \equiv \sum_{u \in B} 1$. The SCDF is discussed more fully in Majure et al. (1996). Now define the quantile function of the SCDF F_B as follows:

$$q(\alpha) \equiv \inf \{z : F_B(z) \geq \alpha\}; \quad \alpha \in [0, 1]. \quad (3.13)$$

In our study, we restrict $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and $B = G$ or L . Note that if $B = G$, then (3.12) and (3.13) are equivalent to the finite-population CDF (2.38) and the finite-population quantile function (2.39), respectively, in Section 2.2.1.

Thus, $F_G(q(\alpha)) = \alpha$, for all α in question, but this is not the case for F_L , since L contains only 9 locations (e.g., $F_L(q(0.1)) = 0.111$). For consistency in our experiment, the quantity to be predicted is always $F_B(q(\alpha))$, rather than α .

Consider the region $B \subset G$, containing $|B|$ locations $\{u_1, \dots, u_{|B|}\}$. Six SCDF predictors are considered in this study.

1. Ordinary Kriging (OK): Define

$$\hat{F}_{B;ok}(z) = \frac{1}{|B|} \sum_{i=1}^{|B|} I(\hat{S}_{ok}(u_i) \leq z); \quad z \in \mathbb{R}, \quad B = G \text{ or } L, \quad (3.14)$$

where $\hat{S}_{ok}(\mathbf{u}_i) = \lambda_{ok}(\mathbf{u}_i)' \mathbf{Z}$ is defined as in (2.13) except that $\mathbf{c}(B)$ is replaced by $\mathbf{c}(\mathbf{u}_i) \equiv (C(\mathbf{u}_i, \mathbf{s}_1), \dots, C(\mathbf{u}_i, \mathbf{s}_n))'$; $i = 1, \dots, |B|$.

2. *Constrained Kriging (CK)*: Define

$$\hat{F}_{B;ck}(z) = \frac{1}{|B|} \sum_{i=1}^{|B|} I(\hat{S}_{ck}(\mathbf{u}_i) \leq z); \quad z \in \mathbb{R}, \quad B = G \text{ or } L, \quad (3.15)$$

where

$$\hat{S}_{ck}(\mathbf{u}_i) = \begin{cases} \lambda_{ck}(\mathbf{u}_i)' \mathbf{Z}, & \text{if } \rho > 0 \\ \bar{Z} + \left\{ \frac{n-(1+\tau^2)}{(1+\tau^2)m(n-m)} \right\}^{\frac{1}{2}} (\sum_{i=1}^m Z(\mathbf{v}_i) - m\bar{Z}), & \text{if } \rho = 0, \end{cases} \quad (3.16)$$

where $\lambda_{ck}(\mathbf{u}_i)' \mathbf{Z}$ is defined as in (2.25) except that $\mathbf{c}(B)$ is replaced by $\mathbf{c}(\mathbf{u}_i) \equiv (C(\mathbf{u}_i, \mathbf{s}_1), \dots, C(\mathbf{u}_i, \mathbf{s}_n))'$; $i = 1, \dots, |B|$. The solution for $\rho = 0$, and an explanation of the terms in (3.16) are given in the Appendix.

3. *Best Predictor (BP)*: From Bayesian decision theory (e.g., Cressie, 1993a, p. 107), the optimal predictor of $F_B(z)$ is $E[F_B(z)|\mathbf{Z}]$. Since we are dealing with Gaussian processes, it is easy to show that

$$E[F_B(z)|\mathbf{Z}] = \frac{1}{|B|} \sum_{i=1}^{|B|} \Phi \left(\frac{y - \hat{S}_{sk}(\mathbf{u}_i)}{\sqrt{\sigma^2 - \mathbf{c}(\mathbf{u}_i)' \Sigma^{-1} \mathbf{c}(\mathbf{u}_i)}} \right); \quad z \in \mathbb{R}, \quad B = G \text{ or } L \quad (3.17)$$

where $\hat{S}_{sk}(\mathbf{u}_i)$ is the simple kriging predictor (2.15) of $S(\mathbf{u}_i)$ and $\Phi(\cdot)$ is the standard normal CDF. Thus we define the best ("plug-in") predictor:

$$\hat{F}_{B;bp}(z) = \frac{1}{|B|} \sum_{i=1}^{|B|} \Phi \left(\frac{y - \hat{S}_{ok}(\mathbf{u}_i)}{\sqrt{\sigma^2 - \mathbf{c}(\mathbf{u}_i)' \Sigma^{-1} \mathbf{c}(\mathbf{u}_i)}} \right); \quad z \in \mathbb{R}, \quad B = G \text{ or } L \quad (3.18)$$

where \hat{S}_{ok} is the ordinary kriging predictor.

4. *Horvitz-Thompson (HT)*: Define

$$\hat{F}_{G;ht}(z) = \frac{1}{n} \sum_{i=1}^n I(Z(\mathbf{s}_i) \leq z); \quad z \in \mathbb{R}, \quad (3.19)$$

where $n = 20$ is the sample size, and

$$\hat{F}_{L;ht}(z) = \begin{cases} \frac{1}{n_L} \sum_{s \in A_L} I(Z(\mathbf{s}) \leq z); & z \in \mathbb{R}, \quad \text{if } n_L > 0 \\ \hat{F}_{G;ht}(z); & z \in \mathbb{R}, \quad \text{if } n_L = 0, \end{cases} \quad (3.20)$$

where $A_L = A \cap L$ and $n_L = |A_L|$.

Note that $\hat{F}_{G;ht}(z)$ is an unbiased estimator of $\sum_{u \in G} I(Z(u) \leq z) / \sum_{u \in G} 1$; $z \in \mathbb{R}$, although this does not account for the measurement error associated with the study (see Section 2.2.5).

5. *Simplified Model (SM)*: Following Fuller (1995) (see Section 2.2.5), we define

$$\hat{F}_{G;sm}(z) \equiv \Phi\left(\frac{z - \hat{\mu}_G}{\hat{\sigma}_G}\right); \quad z \in \mathbb{R}, \quad (3.21)$$

where $\hat{\mu}_G \equiv \bar{Z}$ and $\hat{\sigma}_G^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (Z(s_i) - \bar{Z})^2 - \tau^2$. Also, define

$$\hat{F}_{L;sm}(z) \equiv \begin{cases} \Phi\left(\frac{z - \hat{\mu}_L}{\hat{\sigma}_L}\right); & z \in \mathbb{R}, \text{ if } n_L > 1 \\ \hat{F}_{G;sm}(z); & z \in \mathbb{R}, \text{ if } n_L \leq 1, \end{cases} \quad (3.22)$$

where $\hat{\mu}_L \equiv \bar{Z}_L$, and $\hat{\sigma}_L^2 \equiv \frac{1}{n_L-1} \sum_{u \in A_L} (Z(u) - \bar{Z}_L)^2 - \tau^2$, if $n_L > 1$. In cases where $\hat{\sigma}_B^2$ was negative, a small positive number (viz., $\hat{\sigma}_B^2 = 0.00001$) was substituted, for $B = G$ or L . This predictor is “simplified” in the sense that the assumed model ξ only requires that the marginal distribution of each $S(s)$ be $N(\mu_B, \sigma_B^2)$, $B = G$ or L ; no attempt is made to model the spatial dependence structure.

6. *Simulation Extrapolation Deconvolution (DC)*: Following Stefanski and Bay (1996) (see Section 2.2.5), we define

$$\hat{F}_{G;dc}(z) \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{a}'(D'D)^{-1} D' \mathbf{v}_i; \quad y \in \mathbb{R}, \quad (3.23)$$

where $\mathbf{a}' = (1, -1, 1)$, $D = (1, \eta, \eta^2)$, $\eta = (\eta_1, \dots, \eta_m)'$, $\eta^2 = (\eta_1^2, \dots, \eta_m^2)$, and $\mathbf{v}_i = (\Phi(\frac{z - Z(s_i)}{\tau\sqrt{\eta_1}}), \dots, \Phi(\frac{z - Z(s_i)}{\tau\sqrt{\eta_m}}))'$; $i = 1, \dots, n$, which is equivalent to (2.55). Following the authors' recommendations, we set $\eta = (0.05, 0.2, 0.4, \dots, 2)'$ with $m = 11$.

For the local region L , define

$$\hat{F}_{L;dc}(z) \equiv \frac{1}{n_L} \sum_{i=1}^n \mathbf{a}'(D'D)^{-1} D' \mathbf{v}_i I(s_i \in L); \quad z \in \mathbb{R}, \quad (3.24)$$

provided $n_L \geq 1$. Otherwise $\hat{F}_{L;dc} = \hat{F}_{G;dc}$.

Notice that the first three predictors, ordinary kriging, constrained kriging, and the best predictor are “spatial” predictors in the sense that the invoked model relates directly to the spatial process. The three remaining predictors ignore the spatial structure and can be called “non-spatial” predictors. Observe also that for the three spatial predictors and the Simplified model predictor, the filtering out of measurement error is a straightforward procedure (see Sections 2.1.6 and 2.2.5). In addition, within the geostatistical methodology there exist techniques for estimating the variance of the error process in cases where it is not possible to replicate observations of the study variable at a site (see the discussion about the nugget effect (2.37)). On the other hand, filtering out measurement error adequately for SCDF prediction purposes is a nontrivial problem if no model is invoked for the state variable S .

The best predictor, ordinary kriging, and constrained kriging all require that the spatial covariance parameters be known. The best predictor also requires that $S(\cdot)$ and Z be jointly normal. For the simplified model predictor, it is assumed that each $S(s)$ is $N(\mu, \sigma^2)$, that each $\epsilon(s)$ is $NI(0, \tau^2)$, that $S(\cdot)$ and $\epsilon(\cdot)$ are independent, and that the parameters of these distributions are known. The only modeling assumption of the deconvolution predictor is for the error process to be a zero-mean, white-noise Gaussian process, whose parameter is known. The Horvitz-Thompson estimator has no model assumptions but it does assume that the first-order inclusion probabilities can be calculated; for the probability sampling designs we considered, they are all equal.

3.3 Responses of the Computer Simulation Experiment

The “responses” of this computer simulation experiment are performance criteria of the predictors of the spatial mean and the SCDF. The *design-based* prediction MSE was taken to be the primary criterion, and the design-based prediction bias constituted a secondary criterion. Comparison is in terms of the MSE, subject to the bias not being too large, noting Hájek’s (1971, p. 236) dictum that greatly biased estimators are poor no matter what other properties they have.

Why did we not choose model-based criteria? The spatial models, specified in Section

3.1, were used to ascertain how different analyses, in particular “spatial” vs “non-spatial” analyses, perform under different sampling designs, under different conditions. To use this same spatial model to obtain model-based performance criteria as well, could be perceived as unfairly favoring the “spatial” analyses, so we chose not to do it.

In many cases, expressions for the design-based criteria were unavailable. Consequently, a computer-simulation experiment was conducted such that for all combinations of the experimental factors, 400 realizations of each of the sampling designs were generated and spatial means and SCDFs were predicted from the sampled values in each case. When estimating proportions, such as for the SCDF, 400 realizations guarantees accuracy to the first decimal place.

3.3.1 Spatial-Mean Responses

Suppose $\hat{S}(B; \mathbf{Z})$ is a predictor of $S(B)$, and define $\mathbf{S} \equiv (S(\mathbf{u}_1), \dots, S(\mathbf{u}_{|B|}))'$; $B = G$ or L . Then

$$Bias_{pe}(\hat{S}(B; \mathbf{Z})) \equiv E_{pe}[\hat{S}(B; \mathbf{Z}) - S(B)|\mathbf{S}] \quad (3.25)$$

$$\begin{aligned} MSE_{pe}(\hat{S}(B; \mathbf{Z})) &\equiv E_{pe}[(\hat{S}(B; \mathbf{Z}) - S(B))^2|\mathbf{S}] \\ &= \text{var}_{pe}(\hat{S}(B; \mathbf{Z})|\mathbf{S}) + [Bias_{pe}(\hat{S}(B; \mathbf{Z}))]^2, \end{aligned} \quad (3.26)$$

where $E_{pe}(\cdot|\mathbf{S})$ is the design expectation for the measurement-error model (2.32), (conditional on \mathbf{S}).

The estimators of (3.25) and (3.26) are:

$$\widehat{Bias}_{pe}(\hat{S}(B; \mathbf{Z})) = \frac{1}{400} \sum_{i=1}^{400} [\hat{S}(B; \mathbf{Z}^{(i)}) - S(B)] \quad (3.27)$$

$$\widehat{MSE}_{pe}(\hat{S}(B; \mathbf{Z})) = \frac{1}{400} \sum_{i=1}^{400} [\hat{S}(B; \mathbf{Z}^{(i)}) - S(B)]^2, \quad (3.28)$$

where $\mathbf{Z}^{(i)}$ is the i th random sample; $i = 1, \dots, 400$. Note that $\mathbf{Z}^{(i)}$ will be different from $\mathbf{Z}^{(i')}$ because of both the randomness in the sampling design and the randomness in the measurement error, and that $S(B)$ remains fixed over the 400 pe -realizations.

3.3.2 SCDF Responses

Define F_B as the SCDF of $\{S(s) : s \in B\}$; $B = G$ or L . Let $\hat{F}_B(q(\alpha); \mathbf{Z})$ be a predictor of $F_B(q(\alpha))$; $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Then

$$\text{Bias}_{p\epsilon}(\hat{F}_B(q(\alpha); \mathbf{Z})) = E_{p\epsilon}[\hat{F}_B(q(\alpha); \mathbf{Z}) - F_B(q(\alpha)) | \mathbf{S}] \quad (3.29)$$

$$\text{MSE}_{p\epsilon}(\hat{F}_B(q(\alpha); \mathbf{Z})) = E_{p\epsilon}[(\hat{F}_B(q(\alpha); \mathbf{Z}) - F_B(q(\alpha)))^2 | \mathbf{S}]. \quad (3.30)$$

The estimators of these quantities are:

$$\widehat{\text{Bias}}_{p\epsilon}(\hat{F}_B(q(\alpha); \mathbf{Z})) = \frac{1}{400} \sum_{i=1}^{400} [\hat{F}_B(q(\alpha); \mathbf{Z}^{(i)}) - F_B(q(\alpha))] \quad (3.31)$$

$$\widehat{\text{MSE}}_{p\epsilon}(\hat{F}_B(q(\alpha); \mathbf{Z})) = \frac{1}{400} \sum_{i=1}^{400} [\hat{F}_B(q(\alpha); \mathbf{Z}^{(i)}) - F_B(q(\alpha))]^2, \quad (3.32)$$

where $\mathbf{Z}^{(i)}$ is the i th random sample; $i = 1, \dots, 400$.

3.4 Comments

The simplifications in this study should be noted: First, the variances and covariances of S and Z were completely specified, for the practical reason that variogram-parameter estimation over the 400 p -realizations would have been prohibitive. Thus, a source of variability, which would otherwise occur in practice, has been removed from the spatial predictors, \hat{S}_{ok} and \hat{S}_{ck} . This also has obvious sampling-design implications: There is no longer any need to follow Laslett's (1994) "geostatistical credo" of supplementing a basic grid design with extra clustered points in order to estimate the variogram accurately at short lags. Second, knowledge of the components of $Z(\cdot)$, viz., $S(\cdot)$ and $\epsilon(\cdot)$, easily allowed the kriging equations to be modified so that the measurement-error process $\epsilon(\cdot)$ could be filtered out. In practice, $S(\cdot)$ and $\epsilon(\cdot)$ are usually unobservable individually, and τ^2 is estimated either from replicated observations (preferably) or from the nugget effect of the variogram if some assumptions are made (see Section 2.1.7). The specified error variance τ^2 , was also used in the simplified-model and deconvolution SCDF predictors.

Consider the nonparametric estimator (2.35). In our experiment $E[2\hat{\gamma}(\mathbf{u} - \mathbf{v})] = 2E[(Z(\mathbf{u}) - Z(\mathbf{v}))^2] = 2\gamma(\mathbf{u} - \mathbf{v}) + 2\beta^2(x_u - x_v)^2$; $\mathbf{u} = (x_u, y_u)' \in D_f$, $\mathbf{v} = (x_v, y_v)' \in D_f$, $\beta \in$

$\{0, 1/(8.25)^{-1/2}\}$. Thus, for nonzero β , $2\hat{\gamma}(\mathbf{u} - \mathbf{v})$ has a “leakage” term, $2\beta^2(x_u - x_v)^2$, which will inflate the estimated variogram quadratically, as distances increase in the east-west direction. In our simulation experiment, the leakage term is included in the covariance function, when $\beta \neq 0$ (see Section 2.1.5). That is, we formally used the “variogram”

$$2\gamma^*(\mathbf{u} - \mathbf{v}) \equiv 2(1 - \rho^{\|\mathbf{u} - \mathbf{v}\|}) + 2\beta^2(x_u - x_v)^2; \quad \mathbf{u} = (x_u, y_u)', \mathbf{v} = (x_v, y_v)' \in D_f,$$

from which we obtained the “covariance function”,

$$C^*(\mathbf{u} - \mathbf{v}) \equiv 1 - \gamma^*(\mathbf{u} - \mathbf{v}); \quad \mathbf{u}, \mathbf{v} \in D_f,$$

used for spatial prediction.

The averages of the spatial-mean and the SCDF pe -MSEs (see Section 3.3) over the three S -realizations (see Section 3.2.5) serve as a crude approximation to (2.59), and the three values themselves give some indication of how the pe -MSE *varies* over S -realizations.

4 Results of the Experiment

Not unexpectedly, the design CLU performed extremely poorly throughout the experiment. We did not expect CLU to perform well in this spatial context, but included it because such a design may be used for ecological studies in which a regional process is sampled repeatedly at, or very close to, one (or a few) prespecified spatial location, in the belief that this location is “representative”. This belief can only be supported under the following rather restrictive assumptions on the parameters: β is near zero and ρ is near one. This might occur if the phenomenon being studied “mixes” well (such as the composition of the atmosphere, after several years) but most ecological processes (e.g., timber on forested lands) do not mix well, even after decades. With regard to the study of the nation’s ecological resources, the message from this simulation experiment is clear: No matter how many measurements are taken from so-called “representative sites”, their skill in predicting national and regional ecological resources is extremely low. This is best illustrated by the plots in Figure 4.1, where the spatial mean of the global region is predicted using constrained kriging for varying values of ρ , in the low noise and zero trend case. Shown on the plots is mean squared prediction error (MSE) and absolute

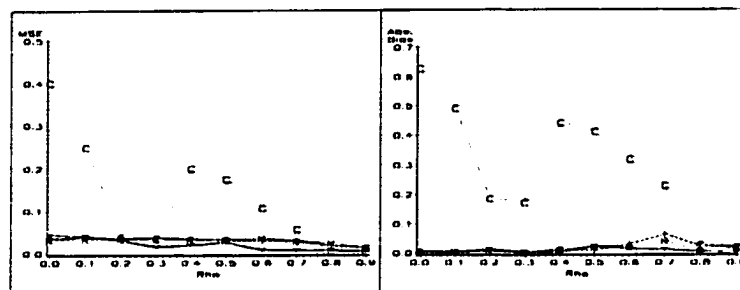


Figure 4.1 Spatial-mean prediction: Mean squared error (MSE) and absolute bias (Abs. Bias) versus spatial-correlation parameter (ρ), respectively. Plots are for constrained kriging (CK), no trend, and low noise over the global region. (Sampling designs: Y=systematic random, T=stratified random, R=simple random, and C=clustered.)

bias for the four designs described in Section 3.2.6, including CLU. The performance of CLU was so poor that one must henceforth doubt the ability of representative sites to say anything about the *regional* behavior of an ecological phenomenon.

In order to present the results of the other three designs on a comparable scale, it was decided to exclude CLU from the rest of this discussion. Not surprisingly, with 20 out of a possible 100 values sampled, spatial-mean prediction over the global region was uniformly good, irrespective of predictor or design, when the response was averaged over all other factors. On the other hand, the results for SCDF prediction over the local region were inconclusive, again not surprisingly, given the small size of the region. Therefore, it was decided to include in this discussion only the results of spatial-mean prediction over the local region, and SCDF prediction over the global region.

4.1 Spatial-Mean Prediction Over the Local Region

The most important features of spatial-mean prediction over the local region L can be summarized from the results presented in Figure 4.2 and Table 4.1.

Figure 4.2 shows MSE and absolute bias of the predictors and designs in a series of plots, corresponding to the levels of the trend and noise factors. All plots are conditioned on the "region" factor being $B = L$ and averaged over those factors not shown. Table 4.1 displays an

Table 4.1 Spatial-mean prediction: Analysis of variance (ANOVA) of mean squared error (MSE) over the local region. Relatively large VR-values are highlighted in bold script. (DF=degrees of freedom, SS=sums of squares, MS=mean squares, and VR=variance ratio, namely, $MS/(\text{Residual MS})$).

Source	DF	SS	MS	VR
TREND	1	121.379031	121.379031	2021.03
REALN	2	23.749024	11.874512	197.72
RHO	9	9.474361	1.052707	17.53
NOISE	1	63.136746	63.136746	1051.26
DESIGN	2	1.456829	0.728415	12.13
PRED	3	251.920640	83.973547	1398.21
TREND*REALN	2	33.274012	16.637006	277.02
TREND*RHO	9	8.080051	0.897783	14.95
TREND*NOISE	1	0.063514	0.063514	1.06
TREND*DESIGN	2	0.446606	0.223303	3.72
TREND*PRED	3	178.165490	59.388497	988.85
REALN*RHO	18	28.758647	1.597703	26.60
REALN*NOISE	2	0.108080	0.054040	0.90
REALN*DESIGN	4	0.190963	0.047741	0.79
REALN*PRED	6	74.958407	12.493068	208.02
RHO*NOISE	9	1.415973	0.157330	2.62
RHO*DESIGN	18	0.567555	0.031531	0.53
RHO*PRED	27	42.102472	1.559351	25.96
NOISE*DESIGN	2	0.042704	0.021352	0.36
NOISE*PRED	3	55.602956	18.534319	308.61
DESIGN*PRED	6	1.793600	0.298933	4.98
TREND*REALN*PRED	6	88.174307	14.695718	244.69
TREND*RHO*PRED	27	31.142113	1.153412	19.20
TREND*NOISE*PRED	3	0.187600	0.062533	1.04
REALN*RHO*PRED	54	44.180658	0.818160	13.62
REALN*NOISE*PRED	6	0.097971	0.016329	0.27
Residual	1213	72.85037	0.06006	
Total	1439	1133.32068		

analysis of variance (ANOVA) of the MSEs of the local (i.e., $B = L$), spatial-mean prediction part of the experiment. The ANOVA here serves merely as an arithmetic partition of selected sums of squares and corresponding variance ratios; no distributional assumptions nor strict statistical inferences are made. The names of the factors in the ANOVA are self-explanatory (e.g., REALN refers to S -realization). The variance ratio, marked "VR" in the table, is the ratio between the "treatment" mean squared error (marked "MS" in the table) and the residual mean squared error.

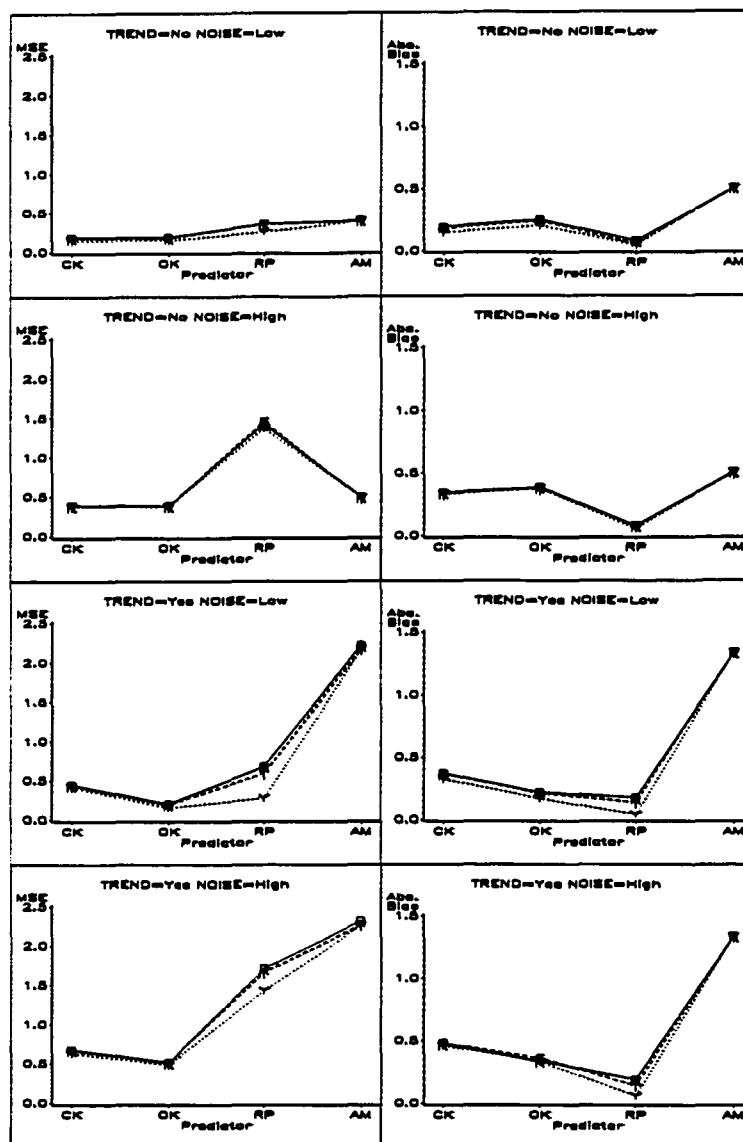


Figure 4.2 Spatial-mean prediction: Mean squared error (MSE) and absolute bias (Abs. Bias) plots for all levels of trend and noise over the local region. On the horizontal axis are the four spatial-mean predictors, constrained kriging (CK), ordinary kriging (OK), regional poststratification (RP), and arithmetic mean (AM). (Sampling designs: Y=systematic random, T=stratified random, and R=simple random.)

The most striking feature in this part of the experiment is the very large predictor effect and the almost negligible design effect (see Figure 4.2 and Table 4.1). We should caution that the small design effect may in part be due to the small size of the global region and the relatively large sample size (i.e., within the constraints of this experiment, the spatial configurations of the three non-clustered designs considered here may not differ too markedly). Nevertheless, in what follows, we see that a spatial analysis is always preferred, regardless of the design. It is also very clear that trend and noise are highly significant factors in this experiment.

Consider the stationary (i.e., no trend) case. Figure 4.2 indicates that ordinary kriging (OK) and constrained kriging (CK) perform best with respect to the MSE criterion, and their bias properties seem to be reasonably good at both noise levels. The arithmetic mean (AM) also performs reasonably well with respect to both criteria, at both noise levels. On the other hand, the regional poststratification (RP) MSE explodes if the noise level is high.

In the presence of trend, AM performs terribly with respect to MSE and bias. The relative precision of OK, CK, and RP do not change much, except that in the presence of trend, OK outperforms CK, but not by much. Notice the slight design effect for RP: here, SYS outperforms the other two designs.

The S -realization effect is not displayed here, but for OK, CK, and RP it is negligible. However, for AM, this effect is large in the presence of trend, accounting for the fairly large variance ratios of the factor REALN and the interactions REALN*PRED and TREND*REALN*PRED in Table 4.1.

The ρ -effect is surprisingly small (see Table 4.1), and a referee has suggested that, because of the geometric decay of the correlation function with distance, values of ρ exceeding 0.9 would also be interesting to look at. Trend and noise appear to be far more important than strength of spatial correlation in the local spatial-mean prediction part of this study.

In conclusion, OK, the spatial BLUP, is the preferred predictor, with CK as a competitive alternative, especially for stationary processes. The arithmetic mean performs very badly in the presence of trend, and RP performs very badly if the data are noisy. This demonstrates that, by using a spatial model describing both large-scale spatial structure (if it exists) and

small-scale spatial structure, observations from *outside* the local region can be used effectively for local spatial-mean prediction.

4.2 SCDF Prediction Over the Global Region

The reader is directed to Figure 4.3, Figure 4.4, Figure 4.5, and Table 4.2 for the following discussion of the results of SCDF prediction over the global region G .

Figure 4.3 displays, on the left, MSE plots of the six predictors for the five α -levels (corresponding to the five quantile-predictands) over both noise levels, for $\beta = 0$. On the right,

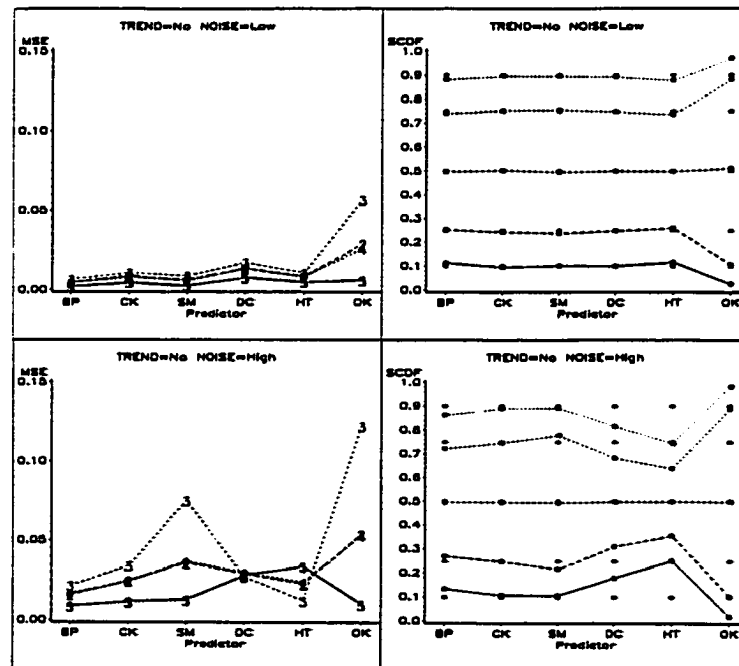


Figure 4.3 Spatial cumulative distribution function (SCDF) Prediction: Mean squared error (MSE) and SCDF-prediction plots for both levels of noise and no trend over the global region. On the horizontal axis are the six SCDF predictors, best predictor (BP), constrained kriging (CK), simplified model (SM), deconvolution (DC), Horvitz-Thompson (HT), and ordinary kriging (OK). The numbers 1, ..., 5 in the MSE plots on the left, represent $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, respectively (e.g., "3" represents median-prediction). In the plots on the right, "*" denotes $F_G(q(\alpha))$; $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, and the dots represent the corresponding predictors.

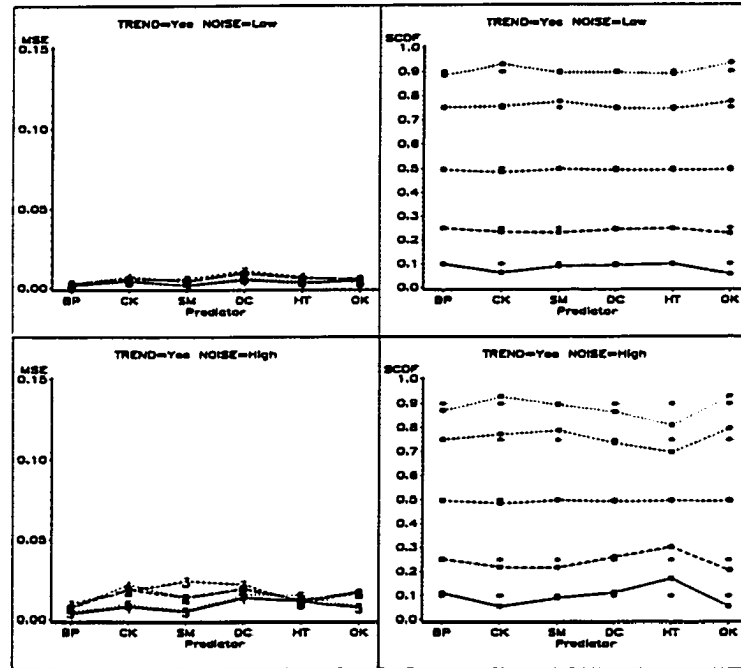


Figure 4.4 Spatial cumulative distribution function (SCDF) Prediction: Mean squared error (MSE) and SCDF-prediction plots for both levels of noise and with trend over the global region. On the horizontal axis are the six SCDF predictors, best predictor (BP), constrained kriging (CK), simplified model (SM), deconvolution (DC), Horvitz-Thompson (HT), and ordinary kriging (OK). The numbers 1, . . . , 5 in the MSE plots on the left, represent $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, respectively (e.g., “3” represents median-prediction). In the plots on the right, “*” denotes $F_G(q(\alpha))$; $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, and the dots represent the corresponding predictors.

the plots display the bias indirectly by showing how the predictors (joined by lines) track their respective predictands (represented by stars). For example, take OK in the top right plot of Figure 4.3: nearly 80% of the probability of the predicted distribution lies within the interquartile range of the “true” distribution. This means that the distribution predicted by OK is too “peaked”, that is, OK yields a surface that is too “smooth” for effective SCDF prediction. All results displayed in Figure 4.3 are for the global (i.e., $B = G$) SCDF part of the experiment, averaged over those factors not shown. Figure 4.4 displays plots similar to those of Figure 4.3, except that here, the trend component is present.

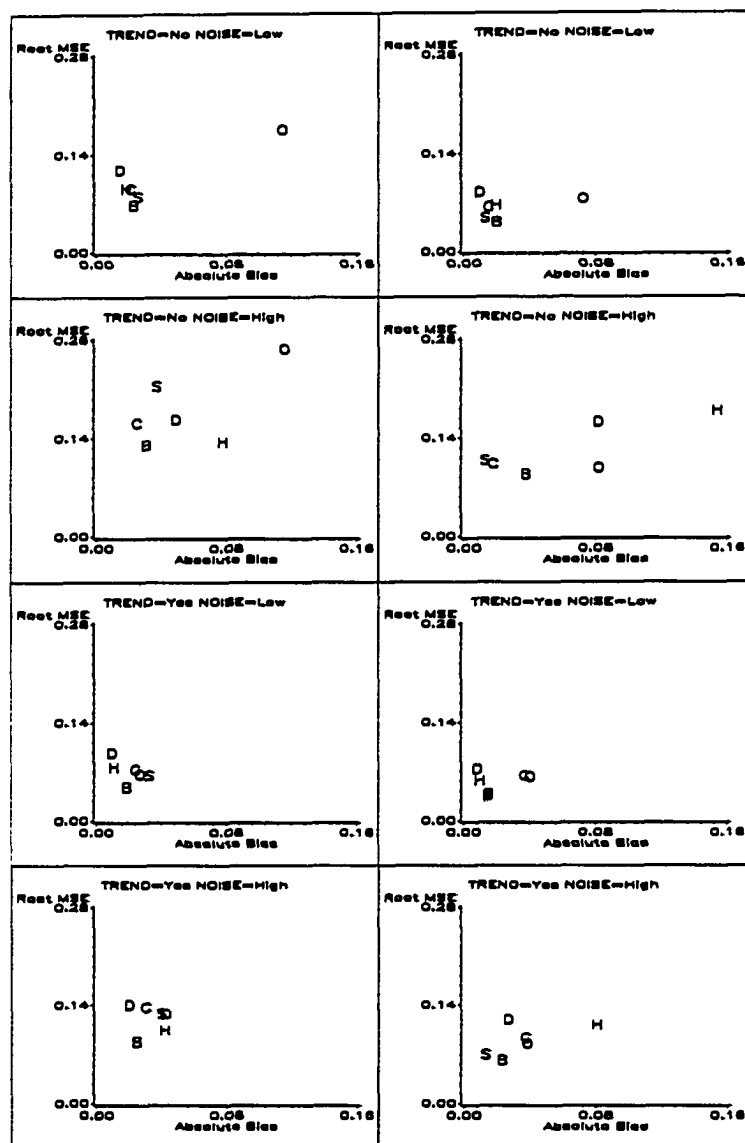


Figure 4.5 Spatial cumulative distribution function (SCDF) Prediction: Plots of root mean squared error (Root MSE) versus absolute bias for all levels of trend and noise over the global region. On the left, root MSE and absolute bias are averaged over the three inner quantiles; on the right, root MSE and absolute bias are averaged over the two outer quantiles. (SCDF predictors: B=best predictor, C=constrained kriging, S=simplified model, D=deconvolution, H=Horvitz-Thompson, and O=ordinary kriging.)

Table 4.2 Spatial cumulative distribution function (SCDF) Prediction: Analysis of variance (ANOVA) of mean squared error (MSE) over the global region. Relatively large VR-values are highlighted in bold script. (DF=degrees of freedom, SS=sums of squares, MS=mean squares, and VR=variance ratio, namely, $MS/(Residual\ MS)$).

Source	DF	SS	MS	VR
TREND	1	0.06014255	0.06014255	10756.48
REALN	2	0.00181021	0.00090510	161.88
RHO	9	0.00334629	0.00037181	66.50
NOISE	1	0.09870255	0.09870255	17652.92
DESIGN	2	0.00104630	0.00052315	93.56
PRED	5	0.05353429	0.01070686	1914.92
TREND*REALN	2	0.00065443	0.00032722	58.52
TREND*RHO	9	0.00242916	0.00026991	48.27
TREND*NOISE	1	0.01533655	0.01533655	2742.94
TREND*DESIGN	2	0.00009405	0.00004702	8.41
TREND*PRED	5	0.03771380	0.00754276	1349.02
REALN*RHO	18	0.00186621	0.00010368	18.54
REALN*NOISE	2	0.00150047	0.00075024	134.18
REALN*DESIGN	4	0.00050638	0.00012659	22.64
REALN*PRED	10	0.00035548	0.00003555	6.36
RHO*NOISE	9	0.00441639	0.00049071	87.76
RHO*DESIGN	18	0.00067312	0.00003740	6.69
RHO*PRED	45	0.01821994	0.00040489	72.41
NOISE*DESIGN	2	0.00000527	0.00000263	0.47
NOISE*PRED	5	0.00765350	0.00153070	273.77
DESIGN*PRED	10	0.00018881	0.00001888	3.38
TREND*REALN*PRED	10	0.00019198	0.00001920	3.43
TREND*RHO*PRED	45	0.01246132	0.00027692	49.53
TREND*NOISE*PRED	5	0.00473842	0.00094768	169.49
REALN*RHO*PRED	90	0.00019311	0.00000215	0.38
REALN*NOISE*PRED	10	0.00039291	0.00003929	7.03
Residual(a)	1837	0.01027120	0.00000559	
Whole-plot Total	2159	0.33844468		
ALPHA	4	0.29125668	0.07281417	1436.54
TREND*ALPHA	5	0.40984807	0.08196961	1617.12
REALN*ALPHA	10	0.00960998	0.00096100	18.96
RHO*ALPHA	45	0.05281559	0.00117368	23.15
NOISE*ALPHA	5	0.54215456	0.10843091	2139.15
ALPHA*DESIGN	10	0.00675377	0.00067538	13.32
ALPHA*PRED	25	0.60924425	0.02436977	480.77
TREND*ALPHA*PRED	25	0.48131501	0.01925260	379.82
REALN*ALPHA*PRED	50	0.00589376	0.00011788	2.33
NOISE*ALPHA*PRED	25	0.17655751	0.00706230	139.33
Residual(b)	8436	0.42759653	0.00005069	
Total	10799	3.35149040		

In Figure 4.5, the square root of the MSE is plotted against absolute bias for each predictor, again in a series of plots corresponding to the levels of trend and noise. Those on the left show results averaged over values corresponding to the three inner-quantile predictions, and those on the right show results averaged over values corresponding to the two outer-quantile predictions.

Table 4.2 is an ANOVA of the prediction MSEs of SCDF predictors over the global region. Its construction is similar to that of Table 4.1, except for the extra subplot factor, ALPHA, which represents the five quantile-predictands under consideration.

Table 4.2 and Figures 4.3 and 4.4 demonstrate the large predictor effect, the small design effect, and the very large effects of the trend and noise factors, echoing the results for the local spatial-mean prediction part of this experiment. In addition, there appears to be a sizable α -effect.

Figures 4.3, 4.4 and 4.5 show the effect of trend on SCDF prediction. In the case of positive trend, all predictors perform fairly comparably. This is not surprising, because if a “mountain” (i.e., a trend) in the data dominates, it will be picked up, irrespective of prediction method! It is in the stationary case, particularly when small-scale variation is dominated by noise, that the merits of the different predictors will likely be demonstrated most clearly.

Thus, considering only the stationary case, a large prediction effect is immediately discernable and, although not displayed in the figures, the design effect is negligible (see Table 4.2). By invoking Hájek’s (1971) dictum that badly biased predictors are unacceptable, it seems that OK fails the bias test (see Figure 4.3). When the noise level is high, the Horvitz-Thompson (HT) predictor, and possibly the deconvolution (DC) predictor, also fail the bias test (see Figure 4.3). Among the others, the best predictor (BP) appears to perform best with respect to the MSE. The CK predictor performs nearly as well and has excellent bias properties. The Simplified model (SM) predictor does not perform well when predicting the middle portion of the distribution when the noise level is high. However, it does predict the tails of the distribution well, irrespective of noise level, accounting in large part for the large variance ratio of the factor ALPHA and the interaction ALPHA*NOISE in Table 4.2.

All predictors behaved consistently over S -realizations, and exhibited little ρ -effect. So,

as in the case of local spatial-mean prediction, trend and noise appear to be more important factors than the strength of spatial correlation in the SCDF-prediction (over G) part of this study (while noting the comment about looking at larger ρ -values given in Section 4.1).

Clearly, the preferred predictor in all cases is BP (see Figures 4.3, 4.4 and 4.5), although it demands the strongest model assumptions (see the discussion on model assumptions, Section 3.2.7), and it may be sensitive to departures from those assumptions. Constrained kriging may be a good alternative if those assumptions cannot be verified to hold at least approximately. Constrained kriging does not require Gaussianity to perform well (Cressie, 1993b); however, it does require that all spatial covariance parameters be known or well estimated. In the low-noise case only, the other predictors, except OK, could be considered acceptable, depending on the number of modeling assumptions one is willing to make.

5 Conclusions

Several conclusions from this study can be drawn but, before we do, we wish to emphasize that these conclusions pertain to the characteristics of the spatial phenomenon described in Section 3.1. In particular, this spatial phenomenon does not contain values of interest that are either clustered or rare. The conclusions follow:

1. Clustered designs, which correspond to so-called “representative-site” selection, should be avoided.
2. Choice of sampling design from among SYS, STS, or SRS designs appears to be unimportant for both spatial-mean and SCDF prediction.
3. For spatial-mean prediction over the local region, the spatial BLUP (i.e., ordinary kriging) is the preferred predictor, although constrained kriging performs competitively, especially for stationary processes. Both predictors require that the spatial covariance parameters be known or well estimated. The regional poststratification predictor should be avoided if the measurement error is large, and the arithmetic mean should be avoided in the presence of a trend component.

4. For SCDF prediction over the global region, the so-called “best predictor” performs best, but requires the strongest model assumptions. Constrained kriging performs well and requires fewer model assumptions. The Simplified model, deconvolution, and Horvitz-Thompson predictors perform well only if the measurement-error component is small. Ordinary kriging should be avoided.
5. Effects of different factors/levels on SCDF prediction are only discernible for larger sample sizes, in comparison to those for spatial-mean prediction. In those cases, 3. and 4. tell us that constrained kriging is a superior predictor.
6. The conclusions stated above were generally consistent across the three S -realizations generated, with minor exceptions as noted in Section 4.

Appendix

The constrained kriging predictor (2.25) is not defined if $m_2 = 0$, and this occurs when $\rho = 0$. We use the limiting result as $\rho \rightarrow 0$ to provide a solution.

Assume that $S(\cdot)$ is a spatial process with constant mean μ and covariance function $C(\mathbf{s}, \mathbf{u})$ and that $Z(\cdot) = S(\cdot) + \epsilon(\cdot)$, where $\epsilon(\cdot)$ is a zero-mean, white-noise process with variance τ^2 . Then, from (2.29), we define the constrained (point) kriging predictor of $S(\mathbf{s}_0)$ as

$$\lambda_{ck}(\mathbf{s}_0)' \mathbf{Z} = \hat{\mu}_{gls} + \left\{ \frac{C(\mathbf{s}_0, \mathbf{s}_0) - \text{var}(\hat{\mu}_{gls})}{\text{var}(\mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu}_{gls} \mathbf{1}))} \right\}^{\frac{1}{2}} \mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu}_{gls} \mathbf{1}).$$

Suppose that exactly $m < n$ sampling locations are equally closest to \mathbf{s}_0 and assume these to be $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$; that is, we have $\|\mathbf{v}_1 - \mathbf{s}_0\| = \dots = \|\mathbf{v}_m - \mathbf{s}_0\| = \min_{j=1, \dots, n} \|\mathbf{s}_j - \mathbf{s}_0\|$.

Thus, for $C(\mathbf{s}, \mathbf{u})$ given by (2.29),

$$\begin{aligned} \mathbf{c}(\mathbf{s}_0) &= \sigma^2 \left(\rho^{\|\mathbf{v}_1 - \mathbf{s}_0\|}, \dots, \rho^{\|\mathbf{v}_1 - \mathbf{s}_0\|}, \rho^{\|\mathbf{v}_{m+1} - \mathbf{s}_0\|}, \dots, \rho^{\|\mathbf{v}_n - \mathbf{s}_0\|} \right)' \\ &= \sigma^2 \rho^{\|\mathbf{v}_1 - \mathbf{s}_0\|} \mathbf{r}', \end{aligned}$$

where $\mathbf{r}' = (1, \dots, 1, \rho^{\|\mathbf{v}_{m+1}-\mathbf{s}_0\|-\|\mathbf{v}_1-\mathbf{s}_0\|}, \dots, \rho^{\|\mathbf{v}_n-\mathbf{s}_0\|-\|\mathbf{v}_1-\mathbf{s}_0\|})$. Consequently,

$$\begin{aligned} \lim_{\rho \rightarrow 0} \lambda_{ck}(\mathbf{s}_0)' \mathbf{Z} &= \lim_{\rho \rightarrow 0} \left\{ \hat{\mu}_{gls} + \left\{ \frac{C(\mathbf{s}_0, \mathbf{s}_0) - \text{var}(\hat{\mu}_{gls})}{\text{var}(\mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu}_{gls} \mathbf{1}))} \right\}^{\frac{1}{2}} \mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu}_{gls} \mathbf{1}) \right\} \\ &= \bar{Z} + \sqrt{\sigma^2 - \text{var}(\bar{Z})} \left(\lim_{\rho \rightarrow 0} \left\{ \frac{\mathbf{r}' \Sigma^{-1} (\mathbf{Z} - \hat{\mu}_{gls} \mathbf{1})}{\sqrt{\text{var}(\mathbf{r}' \Sigma^{-1} (\mathbf{Z} - \hat{\mu}_{gls} \mathbf{1}))}} \right\} \right). \end{aligned}$$

Now, $\lim_{\rho \rightarrow 0} \rho^{\|\mathbf{v}_j-\mathbf{s}_0\|-\|\mathbf{v}_1-\mathbf{s}_0\|} = 0$; $j = m+1, \dots, n$, and so $\lim_{\rho \rightarrow 0} \mathbf{r}' = (\mathbf{1}'_m, 0, \dots, 0)'$. This yields the final result,

$$\lim_{\rho \rightarrow 0} \lambda_{ck}(\mathbf{s}_0)' \mathbf{Z} = \bar{Z} + \left\{ \frac{(n-1)\sigma^2 - \tau^2}{m(n-m)(\sigma^2 + \tau^2)} \right\}^{\frac{1}{2}} \left(\sum_{i=1}^m Z(\mathbf{v}_i) - m\bar{Z} \right). \quad (\text{A.1})$$

Several consequences of (A.1) should be noted:

1. There is no solution if $m = n$. This case is rare and only occurs when $\mathbf{s}_1, \dots, \mathbf{s}_n$ are on the circumference of a circle and \mathbf{s}_0 is at its center. The constrained kriging equations with $\rho > 0$ also break down when this occurs.
2. If $m = 1$ and $\tau^2 = 0$, then $\lim_{\rho \rightarrow 0} \lambda_{ck}(\mathbf{s}_0)' \mathbf{Z} = Z(\mathbf{v}_j)$, where $\|\mathbf{v}_j - \mathbf{s}_0\| = \min_{i=1, \dots, n} \|\mathbf{s}_i - \mathbf{s}_0\|$. That is, constrained kriging yields a piecewise constant prediction surface, constant on Voronoi polygons.

The extension of this result to constrained kriging of blocks is not difficult. Assume that exactly $m \leq n$ sampling locations are equally closest to *any* location in B and assume these to be $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$. Take the i th of these, \mathbf{v}_i , and define $B_i \equiv \{\mathbf{u} \in B : \|\mathbf{u} - \mathbf{v}_i\| = \min_{j=1, \dots, m} \|\mathbf{u} - \mathbf{v}_j\|\}$, $i = 1, \dots, m$. Using similar arguments to the point-prediction case, we obtain

$$\lim_{\rho \rightarrow 0} \lambda_{ck}(B)' \mathbf{Z} = \bar{Z} + \left\{ \frac{n\sigma^2 - |B|(\sigma^2 + \tau^2)}{|B|(\sigma^2 + \tau^2) \{n \sum_{i=1}^m |B_i|^2 - N_m^2\}} \right\}^{\frac{1}{2}} \left(\sum_{i=1}^m |B_i| Z(\mathbf{v}_i) - N_m \bar{Z} \right), \quad (\text{A.2})$$

where $N_m = \sum_{i=1}^m |B_i|$.

Several consequences of (A.2) should be noted:

1. Let A be the set of sampling locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$. If $A \subset B$, then $|B_i| = 1$; $i = 1, \dots, m$, and $m = n$. This means $n \sum_{i=1}^m |B_i|^2 - N_m^2 = n^2 - n^2 = 0$. Hence no solution exists when $B = G$.

2. The equation (A.2) is not defined if $\{n\sigma^2 - |B|(\sigma^2 + \tau^2)\} / \{n \sum_{i=1}^m |B_i|^2 - N_m^2\} < 0$.
3. If $B \equiv s_0$, then (A.2) reduces to (A.1).

References

- Bogardi, I., Bardossy, A., and Duckstein, L. (1985), "Multicriterion network design using geostatistics", *Water Resources Research*, **21**, 199–208.
- Bras, R. L. and Rodriguez-Iturbe, I. (1976), "Network design for the estimation of areal mean of rainfall events", *Water Resources Research*, **12**, 1185–1195.
- Breidt, F. J. (1995a), "Markov chain designs for one-per-stratum sampling", *Survey Methodology*, **21**, 63–70.
- Breidt, F. J. (1995b), "Markov chain designs for one-per-stratum spatial sampling", *Proceedings of the Section on Survey Research Methods*, Vol 1. American Statistical Association, Orlando, FL, 356–361.
- Caselton, W. F. and Zidek, J. V. (1984), "Optimal monitoring network designs", *Statistics & Probability Letters*, **2**, 223–227.
- Cochran, W. G. (1977), *Sampling Techniques*, third ed., Wiley, New York.
- Cook, J. R. and Stefanski, L. A. (1994), "Simulation-extrapolation estimation in parametric measurement error models", *Journal of the American Statistical Association*, **89**, 1314–1328.
- Cordy, C. B. (1993), "An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe", *Probability and Statistics Letters*, **18**, 353–362.
- Cox, D. D., Cox, L. H., and Ensor, K. B. (1995), "Spatial sampling and the environment", Technical Report # 38, National Institute of Statistical Sciences, Research Triangle Park, NC.
- Cressie, N. (1993a), *Statistics for Spatial Data*, rev. ed., Wiley, New York.
- Cressie, N. (1993b), "Aggregation in geostatistical problems", in *Geostatistics Tróia 1992*, Vol. 1, A. Soares (ed.), Kluwer Academic Publishers, Dordrecht, pp. 25–36.
- Cressie, N., Gotway, C. A., and Grondona, M. O. (1990), "Spatial prediction from networks", *Chemometrics and Intelligent Laboratory Systems*, **7**, 251–271.

- Cressie, N. and Ver Hoef, J. M. (1993), "Spatial statistical analysis of environmental and ecological data", in *Environmental Modeling with GIS*, M. F. Goodchild, B. O. Parks, and L. T. Steyaert (eds.), Oxford University Press, New York, pp. 404-413.
- Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments", *Journal of the American Statistical Association*, **86**, 953-963.
- Dalenius, T., Hájek, J., and Zubrzycki, S. (1960), "On plane sampling and related geometrical problems", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, CA, 125-150.
- De Gruijter, J. J. and ter Braak, C. J. F. (1990), "Model-free estimation from spatial samples: A reappraisal of classical sampling theory", *Mathematical Geology*, **22**, 407-415.
- Englund, E. J. and Heravi, N. (1993), "Conditional simulation: Practical application for sampling design optimization", in *Geostatistics Tróia 1992*, Vol. 2, A. Soares (ed.), Kluwer Academic Publishers, Dordrecht, pp. 612-624.
- Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, third ed., Oliver & Boyd, Edinburgh.
- Fuller, W. A. (1995), "Estimation in the presence of measurement error", *International Statistical Review*, **6**, 121-147.
- Goldberger, A. S. (1962), "Best linear unbiased prediction in the generalized linear regression model", *Journal of the American Statistical Association*, **57**, 369-375.
- Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Wadsworth & Brooks/Cole. Pacific Grove, CA.
- Guttorp, P., Le, N. D., Sampson P. D., and Zidek, J. V. (1993), "Using entropy in the redesign of an environmental monitoring network", in *Multivariate Environmental Statistics*. G. P. Patil and C. R. Rao (eds.), North Holland, pp. 347-386.
- Haas, T. C. (1992), "Redesigning continental-scale monitoring networks", *Atmospheric Environment*, **26A**, 3323-3333.
- Hájek J. (1971), "Comment on paper by D. Basu", in *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.), Holt, Rinehart and Winston, Toronto, p. 236.
- Hammersley, J. M. and Handscomb, D. C. (1964), *Monte Carlo Methods*, Methuen, London.
- Horvitz, D. G. and Thompson, D. F., (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, **47**, 663-685.

- Isaki, C. T. and Fuller, W. A. (1982), "Survey design under the regression superpopulation model", *Journal of the American Statistical Association*, **77**, 89-96.
- Laslett, G. M. (1994), "Kriging and splines: An empirical comparison of their predictive performance in some applications", *Journal of the American Statistical Association*, **89**, 391-409.
- Majure, J. J., Cook, D., Cressie, N., Kaiser, M., Lahiri, S., and Symanzik, J. (1996), "Spatial CDF estimation and visualization with applications to forest health monitoring", *Computer Science and Statistics*, **27**, 93-101.
- McArthur, R. D. (1987), "An evaluation of sample designs for estimating a locally concentrated pollutant", *Communications in Statistics. Simulation and Computation*, **16**, 735-759.
- Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996), "A semiparametric transformation approach to estimating usual daily intake distributions", *Journal of the American Statistical Association*, **91**, 1440-1449.
- Olea, R. A. (1984), "Sampling design optimization for spatial functions", *Mathematical Geology*, **16**, 369-392.
- Overton, W. S. (1993), "Probability sampling and population inference in monitoring programs", in *Environmental Modeling with GIS*, M. F. Goodchild, B. O. Parks, and C. T. Steyaert (eds.), Oxford University Press, New York, pp. 470-480.
- Overton, W. S. and Stehman, S. V. (1993), "Properties of designs for sampling continuous spatial resources from a triangular grid", *Communications in Statistics. Theory and Methods*, **22**, 2641-2660.
- Pérez-Abreu, V. and Rodríguez, J. E. (1996), "Index of effectiveness of a multivariate environmental monitoring network". *Environmetrics*, **7**, 489-501.
- Särndal C-E., Swensson B., and Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Seber, G. A. F. and Thompson, S. K. (1994), "Environmental adaptive sampling", in *Handbook of Statistics*, Vol. 12, G. P. Patil and C. R. Rao (eds.), Elsevier Science B.V., pp. 201-220.
- Stefanski, L. A. and Bay, J. M. (1996), "Simulation extrapolation deconvolution of finite population cumulative distribution function estimators", *Biometrika*, **83**, 406-417.
- Stevens, D. L. (1994), "Implementation of a national monitoring program", *Journal of Environmental Management*, **42**, 1-29.

Thompson, S. K. (1992), *Sampling*, Wiley, New York.

Wolter, K. M. (1984), "An investigation of some estimators of variance for systematic sampling", *Journal of the American Statistical Association*, **79**, 781–790.

Zeleny, M. (1982), *Multiple Criteria Decision Making*, McGraw-Hill, New York.

Acknowledgements

The research presented in this article was supported by the U. S. Environmental Protection Agency under co-operative agreement #CR822919-01-0. The article has not been subjected to the review of the EPA and thus does not reflect the view of the agency and no official endorsement should be inferred. The authors would like to express their appreciation to Scott Overton and an anonymous referee for their helpful comments.

THE PREDICTION OF MEASURABLE FUNCTIONALS OF SPATIAL PROCESSES USING DATA CONTAMINATED WITH MEASUREMENT ERROR

A paper, a portion of which is to be submitted to the Journal of the American Statistical Association

Jeremy Aldworth and Noel Cressie

Abstract

In this paper, a spatial model that explicitly includes a measurement-error component is proposed, to accommodate the fact that data is almost always contaminated with measurement error. Linear predictors can easily accommodate this measurement-error component, but this is not true of nonlinear predictors, which may be substantially biased if the measurement-error variance is large. For the prediction of nonlinear functionals of spatial processes, constrained kriging is examined in detail, especially with regard to its existence conditions, its geometric interpretation, its applicability to certain “nonspatial” problems, and its relationship with conditional simulation. The theory supporting constrained kriging is extended to the “covariance-matching” case where multiple predictions are required simultaneously.

1 Introduction

The motivation for the research presented in this paper comes from a problem commonly encountered in the arena of environmental regulation. For example, the U. S. Environmental Protection Agency (EPA) has a federal mandate (e.g., Sections 108 and 109, Clean Air Act, 1970) to identify air pollutants injurious to the public health, and to propose “standards” (i.e., threshold levels of the pollutants) which, if exceeded, may endanger the public health. A major

component of air pollution is particulate matter (PM), and this is categorized by size. For example, PM_{10} refers to the category of particulate matter in which particles do not exceed $10\ \mu\text{m}$ in aerodynamic diameter, in a cubic meter of air. Aerodynamic diameter is defined as the diameter of a spherical particle with equal settling velocity but a material density of $1\text{g}/\text{cm}^3$, normalizing particles of different shapes and densities (U. S. EPA, 1996a; p. 3-8). The PM_{10} measurement is an indicator for particles that penetrate to the tracheobronchial and gas-exchange regions of the lung (U. S. EPA, 1996b, p. IV-3a). The EPA recommends adopting an annual PM_{10} standard in the range 40 to $50\ \mu\text{g}/\text{m}^3$ and a 24-hour standard of $150\ \mu\text{g}/\text{m}^3$. This means that if the PM_{10} -level of some region (e.g., the city of Pittsburgh, PA) exceeds the standard for the appropriate time period, then that region is deemed to be out of compliance with federal regulations. The problem here is how to predict reliably whether or not the PM_{10} -level exceeds the standard, based on data contaminated with measurement error, that are obtained from a small number of monitoring stations in the region of interest. And, even if the entire region is in PM_{10} compliance, there may be local neighborhoods that are not in compliance (e.g., neighborhoods downwind of pollution-emitting factories). It may be difficult to identify such neighborhoods based on the data available, but it may be possible to predict the *proportion* of the entire region not in compliance. Such predictands are clearly *nonlinear* functions of the study variable, here the PM_{10} reading as a function of location. More generally, how do we predict reliably *any* measurable function of some study variable, based on data contaminated with measurement error? We attempt to address problems of this nature in this paper.

Consider some study variable(s) that varies over some spatial domain (e.g., rainfall over Iowa, PM_{10} -levels over Pittsburgh, or piezometric head in the Wolfcamp Aquifer of SW Texas), and suppose that measurements associated with it (them) are taken at selected locations in the domain. We are interested in the prediction of some measurable function of the study variable, based on the observed data, but it is unlikely that the study variable can be measured exactly, (i.e., without measurement error). Notice that in much of the geostatistics literature, there is an implicit assumption that observations are exact measurements of the study variable in

question (e.g., Journel and Huijbregts, 1978; Rivoirard, 1994). In this paper, we make no such assumption; indeed, we propose a spatial model that explicitly includes a measurement-error component and this is presented in Section 2.

Much attention has been given to the prediction of *linear* functions of the study variable (i.e., kriging), such as prediction at some unsampled location or the prediction of the average over some subregion in the domain (e.g., Journel and Huijbregts, 1978; Cressie, 1993a). It has been shown by, for example, Gotway and Cressie (1993), that the ordinary-kriging and universal-kriging prediction methodology has useful optimality properties for predictands of this kind, particularly if the distributional characteristics of the study variable (as seen in the data) appear to be Gaussian. Ordinary-kriging and universal-kriging predictors are linear in the data and are simple to construct (e.g., Journel and Huijbregts, 1978, Ch. V; Cressie, 1993a, Ch. 3), and this prediction methodology can be extended easily to the case where the data are multivariate (e.g., Myers, 1982; Ver Hoef and Cressie, 1993). However, if the data suggest that a non Gaussian distribution is appropriate for the study variable, then these linear kriging predictors may be far from optimal. In such a case, alternative predictors, nonlinear in the data, such as indicator kriging, indicator cokriging, and disjunctive kriging, have been proposed as alternatives to the linear kriging predictors (e.g., Journel and Huijbregts, 1978, Ch. VIII; Cressie, 1993a, Sect. 5.1). In the formulation of these nonlinear kriging predictors, it has been assumed that no measurement error exists, and these predictors may perform poorly in the presence of substantial measurement error. All these predictors, linear and nonlinear, are discussed in Section 3.

The problem of the prediction of *nonlinear* functions of the study variable (Gaussian or not) is more difficult. For example, one may want to predict whether the study variable is in compliance or not, or to predict the proportion of a region that is out of compliance; both involve indicator functions that yield nonlinear functions. For such problems, the linear kriging predictors (i.e., ordinary and universal kriging) generally perform poorly because they are usually too smooth (e.g., Journel and Huijbregts, 1978, Ch. VIII). Nonlinear kriging predictors (i.e., indicator kriging, indicator cokriging, and disjunctive kriging) may perform better than

the linear kriging predictors but possibly only if the measurement error is not substantial; versions of them that account for measurement error have yet to be developed. In this paper, we propose an alternative predictor, based on linear kriging methodology, but with a constraint forcing the variance of the predictor to match that of the predictand. This predictor, called “constrained kriging” (Cressie, 1993b), has useful unbiasedness properties, approximate second-order optimality properties, and it easily filters out measurement error. These topics are discussed in Section 4.

The theory supporting constrained kriging is extended to the “covariance-matching” case where multiple predictions are required simultaneously, and this is presented in Section 5. In Section 7 a study of PM_{10} data in the Pittsburgh area is presented where these predictors are applied. Finally, Section 8 contains discussion and conclusions.

2 Measurement-error Spatial Model

A *spatial process* is a real- (or vector-) valued stochastic process

$$Z(\cdot) \equiv \{Z(s) : s \in D\}, \quad (2.1)$$

where $D \subset \mathbb{R}^d$. In this paper, unless otherwise stated (e.g., Section 3.6), we consider only real-valued $Z(\cdot)$.

Suppose that *data* from the spatial process $Z(\cdot)$ consist of the observations $\{Z(s_1), \dots, Z(s_n)\}$, taken at the locations in $A \equiv \{s_1, \dots, s_n\} \subset D$, where A is called a *sample*. For our purposes, we assume that the sample has been specified somehow and so it is fixed. Readers interested in how randomness in A can be considered within the context of geostatistics might consult the review given in Aldworth and Cressie (1998).

The observations are assumed to be contaminated with measurement error, for which the following model is useful:

$$Z(s) = S(s) + \epsilon(s); \quad s \in D, \quad (2.2)$$

where $S(\cdot)$ represents the “signal” component and $\epsilon(\cdot)$ represents the “noise” component of

$Z(\cdot)$. The signal process is further decomposed as

$$S(s) = \mu(s) + \delta(s); \quad s \in D, \quad (2.3)$$

where $\mu(\cdot)$ is the large-scale, deterministic, mean structure of the process (i.e., trend), and $\delta(\cdot)$ is the small-scale stochastic structure that models the spatial dependence among the data. That is,

$$E[\delta(s)] = 0; \quad s \in D \quad (2.4)$$

$$\text{cov}[\delta(s), \delta(u)] \equiv C(s, u); \quad s, u \in D, \quad (2.5)$$

where $C(\cdot, \cdot)$ is the *covariance function* of $\delta(\cdot)$ (i.e., of $S(\cdot)$), and $\epsilon(\cdot)$ represents a measurement-error process assumed to be independent of $S(\cdot)$. Measuring instruments may be biased, or they may yield correlated errors but, in this paper, we assume a simplified zero-mean, white-noise structure for the ϵ -process and denote

$$\text{var}(\epsilon(s)) \equiv \tau^2; \quad s \in D. \quad (2.6)$$

Define the covariance function of $Z(\cdot)$ as

$$C_z(s, u) \equiv \text{cov}[Z(s), Z(u)]; \quad s, u \in D; \quad (2.7)$$

then

$$C_z(s, u) = \begin{cases} C(s, u) + \tau^2, & s = u \\ C(s, u), & s \neq u. \end{cases} \quad (2.8)$$

Another useful measure of the spatial dependence that is found in $Z(\cdot)$ is the *variogram*, defined as

$$\begin{aligned} 2\gamma_z(s, u) &\equiv \text{var}[Z(s) - Z(u)] \\ &= C_z(s, s) + C_z(u, u) - 2C_z(s, u); \quad s, u \in D \\ &= C(s, s) + C(u, u) - 2C(s, u) + 2\tau^2; \quad s \neq u \\ &\equiv 2(\gamma(s, u) + \tau^2); \quad s \neq u, \end{aligned} \quad (2.9)$$

where $2\gamma(\cdot, \cdot)$ is the variogram of $\delta(\cdot)$ (i.e., of $S(\cdot)$). When $s = u$, $2\gamma_z(s, u) = 2\gamma(s, u) = 0$. The quantity $\gamma_z(\cdot, \cdot)$ is called the *semivariogram* of $Z(\cdot)$. Observe that if $\tau^2 = 0$, then $C(\cdot, \cdot) \equiv C_z(\cdot, \cdot)$ and $\gamma(\cdot, \cdot) \equiv \gamma_z(\cdot, \cdot)$.

It is very important to realize that if the measurement-error model (2.2) is assumed, then we are interested in predicting some functional of the “noiseless” S -process over D , but what we actually observe are “noisy” data $\{Z(s_1), \dots, Z(s_n)\}$. For example, point prediction involves predicting $S(s_0)$ based on data $\{Z(s_1), \dots, Z(s_n)\}$, where $s_0 \in D$ is a given point location.

Replicated Observations

Suppose that replicate observations are taken at one or more of the sites in A . Then measurement-error spatial model (2.2) can be generalized to accommodate the replicate observations as follows,

$$Z_i(s_j) = S(s_j) + \epsilon_i(s_j); \quad i = 1, \dots, n_j, \quad s_j \in D, \quad (2.10)$$

where, for each $i = 1, \dots, n_j$, $\epsilon_i(\cdot)$ is an independent, zero-mean, white-noise process with

$$\text{var}(\epsilon_i(s_j)) \equiv \tau^2; \quad i = 1, \dots, n_j, \quad s_j \in D.$$

Thus, if $n_j > 1$ observations are taken at site $s_j \in D$, then $Z_i(s_j)$ and $Z_{i'}(s_j)$ differ from one another only in their error terms, $\epsilon_i(s_j)$ and $\epsilon_{i'}(s_j)$, respectively, for $i \neq i'$. Then,

$$\begin{aligned} \text{cov}[Z_i(s_j), Z_{i'}(s_{j'})] &= \text{cov}[S_i(s_j) + \epsilon_i(s_j), S_{i'}(s_{j'}) + \epsilon_{i'}(s_{j'})] \\ &= \begin{cases} C(s_j, s_j) + \tau^2; & i = i', \quad j = j', \\ C(s_j, s_{j'}); & \text{otherwise.} \end{cases} \end{aligned}$$

Stationarity and Isotropy

The process $S(\cdot)$ is said to be *first-order stationary* if

$$E[S(s)] \equiv \mu(s) = \mu; \quad s \in D. \quad (2.11)$$

If $S(\cdot)$ is first-order stationary, and

$$\text{cov}[S(s), S(u)] \equiv C(s, u) = C^o(h); \quad s, u \in D, \quad (2.12)$$

where $\mathbf{h} = \mathbf{s} - \mathbf{u}$ (i.e., $C^o(\cdot)$ is a function only of the vector difference \mathbf{h}), then $S(\cdot)$ is said to be *second-order stationary*. If, in addition,

$$C^o(\mathbf{h}) = C^\#(||\mathbf{h}||), \quad (2.13)$$

that is, $C^o(\mathbf{h})$ is in fact a function only of the *length* of \mathbf{h} (irrespective of direction), then $S(\cdot)$ is said to be *isotropic*.

If $S(\cdot)$ is first-order stationary, and

$$\text{var}[S(\mathbf{s}) - S(\mathbf{u})] \equiv 2\gamma(\mathbf{s}, \mathbf{u}) = 2\gamma^o(\mathbf{h}); \quad \mathbf{s}, \mathbf{u} \in D, \quad (2.14)$$

where $\mathbf{h} = \mathbf{s} - \mathbf{u}$ (i.e., $\gamma^o(\cdot)$ is a function only of the vector difference \mathbf{h}), then $S(\cdot)$ is said to be *intrinsically stationary*. If $S(\cdot)$ is second-order stationary, then $S(\cdot)$ is also intrinsically stationary, although the converse is not true. That is, intrinsic stationarity is a more general type of stationarity than the usual second-order stationarity.

3 Prediction of Linear Functionals of the S -process

Prediction presupposes some target. If the measurement-error spatial model is assumed, then the target, or predictand, will be some functional of $S(\cdot)$ (i.e., not some functional of $Z(\cdot)$). In this section we consider only *linear* predictands, examples of which are given below.

The *point predictand* at location $\mathbf{s}_0 \in D$ is defined as

$$h(S(\cdot)) = S(\mathbf{s}_0). \quad (3.1)$$

The *block predictand* (also known as the *spatial mean*) for block $B \subset D$, of nonzero volume, is defined as

$$\begin{aligned} h(S(\cdot)) &= S(B) \\ &\equiv \frac{1}{|B|} \int_B S(\mathbf{s}) \, d\mathbf{s}, \end{aligned} \quad (3.2)$$

where $|B| = \int_B 1 \, d\mathbf{s} > 0$.

Cartier's Formula

Suppose that $B \subset D$ has nonzero volume, and let \mathbf{x} be a point chosen at random within B (i.e., the random location \mathbf{x} is uniformly distributed in B). Then,

$$\begin{aligned}
 E[S(\mathbf{x})|S(B)] &= E\{E[S(\mathbf{x})|S(B)]|\mathbf{x}=\mathbf{s}\} \\
 &= \frac{1}{|B|} \int_B E[S(\mathbf{s})|S(B)] d\mathbf{s} \\
 &= \frac{1}{|B|} \int_B S(\mathbf{s}) d\mathbf{s} \\
 &= S(B).
 \end{aligned} \tag{3.3}$$

Equation (3.3) is known as Cartier's formula, and Lantuejoul (1988) notes several consequences of this.

First,

$$E[S(B)] = E\{E[S(\mathbf{x})|S(B)]\} = E[S(\mathbf{x})].$$

Therefore, if $S(\cdot)$ is first-order stationary (see Section 2), then

$$E[S(B)] = E[S(\mathbf{x})] = E[S(\mathbf{s})]; \quad \mathbf{s} \in D, \tag{3.4}$$

which is a result that could have been derived directly.

Second, for any convex function g ,

$$g(S(B)) \leq E[g(S(\mathbf{x})|S(B))],$$

and hence,

$$E[g(S(B))] \leq E\{E[g(S(\mathbf{x})|S(B))]\} = E[g(S(\mathbf{x}))].$$

Thus,

$$E[(S(B))^2] \leq E[(S(\mathbf{x}))^2],$$

since $g(\cdot) \equiv (\cdot)^2$ is a convex function. Therefore, if $\text{var}(S(\mathbf{s}))$ does not depend on \mathbf{s} (which is implied by second-order stationarity), then

$$E[(S(B))^2] \leq E[(S(\mathbf{x}))^2] = E[(S(\mathbf{s}))^2]; \quad \mathbf{s} \in D,$$

that is,

$$\text{var}(S(B)) \leq \text{var}(S(\mathbf{s})); \quad \mathbf{s} \in D. \tag{3.5}$$

3.1 Best Predictor

Suppose we wish to predict $S(B)$; $B \in D$. (Note that B may consist of the single element $\{s_0\}$.) From Bayesian decision theory (e.g. Cressie, 1993a, p. 107), the *best predictor* of $S(B)$ (with respect to squared-error loss) is

$$\hat{S}_{bp}(B) \equiv E[S(B)|\mathbf{Z}], \quad (3.6)$$

where $\mathbf{Z} \equiv (Z(s_1), \dots, Z(s_n))'$ is the vector of data, and $A \equiv \{s_1, \dots, s_n\} \subset D$ are the locations at which observations are taken.

If $(S(\cdot), \epsilon(\cdot))$ is a bivariate Gaussian process, then $\hat{S}_{bp}(B)$ is *linear* in \mathbf{Z} , and it has a simple form (see Section 3.2.1).

If $(S(\cdot), \epsilon(\cdot))$ is not a bivariate Gaussian process, then $\hat{S}_{bp}(B)$ is typically *nonlinear* in \mathbf{Z} and may be difficult to obtain. In such cases, nonlinear predictors (i.e., predictors that are nonlinear in \mathbf{Z}) have been formulated to approximate $\hat{S}_{bp}(B)$ (see Section 3.7).

3.2 Linear Predictors

3.2.1 Best (Heterogeneous) Linear Predictor

The *best (heterogeneous) linear predictor*, also known as the *simple kriging predictor*, of $S(B)$ is given by

$$\hat{S}_{sk}(B) \equiv \mu(B) + \mathbf{c}(B)' \Sigma^{-1}(\mathbf{Z} - \boldsymbol{\mu}), \quad (3.7)$$

where $\mu(B) \equiv E(S(B))$, $\boldsymbol{\mu} \equiv E(\mathbf{Z})$, $\Sigma \equiv \text{var}(\mathbf{Z}) = \text{var}((S(s_1), \dots, S(s_n))') + \tau^2 \mathbf{I}$, \mathbf{I} is the $n \times n$ identity matrix, $\mathbf{c}(B) \equiv \text{var}(\mathbf{Z}, S(B)) = (C(s_1, B), \dots, C(s_n, B))'$, and $C(s, B) \equiv \frac{1}{|B|} \int_B C(s, u) du$; if $B = \{s_0\}$ then $\mathbf{c}(s_0) \equiv \text{var}(\mathbf{Z}, S(s_0)) = (C(s_1, s_0), \dots, C(s_n, s_0))'$. The result is not hard to prove and can be found, for example, in Gotway and Cressie (1993). Note that only Σ is a function of τ^2 .

The mean-squared prediction error (MSPE) of $\hat{S}_{sk}(B)$ is given by

$$\begin{aligned} \text{MSPE}(\hat{S}_{sk}(B)) &\equiv E[(\hat{S}_{sk}(B) - S(B))^2] \\ &= \text{var}(\hat{S}_{sk}(B) - S(B)) \\ &= C(B, B) - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B), \end{aligned} \quad (3.8)$$

where $C(B, B) \equiv \text{var}(S(B)) = |B|^{-2} \int_B \int_B C(s, u) ds du$, if $|B| > 0$; otherwise, $C(s_0, s_0) \equiv \text{var}(S(s_0))$.

If $(S(\cdot), \epsilon(\cdot))$ is a Gaussian process, then

$$\hat{S}_{sk}(B) = \hat{S}_{bp}(B); \quad (3.9)$$

this is shown, for example, by Graybill (1976, Section 12.2).

Suppose that the parametric form of both $\mu(\cdot)$ and $C(\cdot, \cdot)$ are known, and that $\mu(\cdot)$ and $C(\cdot, \cdot)$ are governed by the parameter vectors, β and θ , respectively; that is, $\mu(\cdot) \equiv \mu(\cdot; \beta)$ and $C(\cdot, \cdot) \equiv C(\cdot, \cdot; \theta)$. Then $\hat{S}_{sk}(B)$ is a function of the (typically unknown) parameters $\{\beta, \theta, \tau^2\}$.

3.2.2 Best Linear Unbiased Predictor

Suppose that β is unknown, and that $\mu(\cdot; \beta)$ is linear in β ; that is,

$$\mu(s; \beta) = \sum_{j=1}^{p+1} x_{j-1}(s) \beta_{j-1}, \quad (3.10)$$

where $\beta = (\beta_0, \dots, \beta_p)'$ is the vector of unknown parameters, and $\{x_0(\cdot), \dots, x_p(\cdot)\}$ are known functions. Upon integrating both sides, equation (3.10) can be written as

$$\mu(B; \beta) = \mathbf{x}(B)' \beta; \quad B \subset D. \quad (3.11)$$

where $\mathbf{x}(B) \equiv (x_0(B), \dots, x_p(B))'$, and $x_j(B) \equiv \frac{1}{|B|} \int_B x_j(u) du$; $j = 1, \dots, p$. If $B = \{s_0\}$, then (3.10) is equivalent to (3.11) with $\mathbf{x}(s_0) \equiv (x_0(s_0), \dots, x_p(s_0))'$.

From linear-model theory, we have that the generalized-least-squares estimator of β is also the *best linear unbiased estimator* (BLUE) of β , and is defined as

$$\hat{\beta} \equiv (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Z}, \quad (3.12)$$

where $\mathbf{X} = (\mathbf{x}(s_1), \dots, \mathbf{x}(s_n))'$, $\mathbf{x}(s_i) \equiv (x_0(s_i), \dots, x_p(s_i))'$; $i = 1, \dots, n$, and $\Sigma = \text{var}(\mathbf{Z})$.

Cressie(1993a, Section 3.4) shows that the *best linear unbiased predictor* (BLUP), also known as the *universal kriging predictor*, of $S(B)$ is defined as

$$\hat{S}_{uk}(B) \equiv \mathbf{x}(B)' \hat{\beta} + \mathbf{c}(B)' \Sigma^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}), \quad (3.13)$$

where the terms in this equation are defined as in equations (3.10), (3.11), and (3.12).

The MSPE of $\hat{S}_{uk}(B)$ is given by

$$\begin{aligned}
 \text{MSPE}(\hat{S}_{uk}(B)) &\equiv E[(\hat{S}_{uk}(B) - S(B))^2] \\
 &= C(B, B) - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \\
 &\quad + (\mathbf{x}(B)' - \mathbf{c}(B)' \Sigma^{-1} \mathbf{X})(\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{x}(B) - \mathbf{X}' \Sigma^{-1} \mathbf{c}(B)) \quad (3.14) \\
 &= \text{MSPE}(\hat{S}_{sk}(B)) \\
 &\quad + (\mathbf{x}(B)' - \mathbf{c}(B)' \Sigma^{-1} \mathbf{X})(\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{x}(B) - \mathbf{X}' \Sigma^{-1} \mathbf{c}(B)) \\
 &\geq \text{MSPE}(\hat{S}_{sk}(B)),
 \end{aligned}$$

since Σ is non-negative definite (n.n.d.). Observe that if $\text{var}(\hat{\beta}) = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1}$ is “small” in the sense that its largest eigenvalue is small, then $\text{MSPE}(\hat{S}_{uk}(B)) \simeq \text{MSPE}(\hat{S}_{sk}(B))$.

If $S(\cdot)$ is first-order stationary (i.e., $\mu(\mathbf{s}) \equiv \mu$; $\mathbf{s} \in D$), then the BLUP of $S(B)$ is called the *ordinary kriging predictor*, and is written

$$\hat{S}_{ok}(B) \equiv \hat{\mu} + \mathbf{c}(B)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu} \mathbf{1}), \quad (3.15)$$

where $\hat{\mu} \equiv \mathbf{1}' \Sigma^{-1} \mathbf{Z} / \mathbf{1}' \Sigma^{-1} \mathbf{1}$ is the BLUE of μ , and $\mathbf{1}$ is a vector of n ones.

The MSPE of $\hat{S}_{ok}(B)$ is given by

$$\begin{aligned}
 \text{MSPE}(\hat{S}_{ok}(B)) &= C(B, B) - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) + (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} (1 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{1})^2 \quad (3.16) \\
 &= \text{MSPE}(\hat{S}_{sk}(B)) + (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} (1 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{1})^2 \\
 &\geq \text{MSPE}(\hat{S}_{sk}(B)).
 \end{aligned}$$

Clearly, if $\text{var}(\hat{\mu}) = (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1}$ is small, then $\text{MSPE}(\hat{S}_{ok}(B)) \simeq \text{MSPE}(\hat{S}_{sk}(B))$.

If $(S(\cdot), \epsilon(\cdot))$ is a Gaussian process, and $\text{var}(\hat{\beta})$ (or $\text{var}(\hat{\mu})$) is small, then clearly \hat{S}_{uk} (or \hat{S}_{ok}) cannot be far from the best (with respect to squared-error loss) predictor.

Usually, the parameters $\{\theta, \tau^2\}$ are also unknown, but are estimated and “plugged in” to the prediction equations as if they were the “true” parameter values (see Cressie, 1993a, Sections 2.4 and 2.6). This means that the uncertainty in estimating θ and τ^2 , if unknown, is not reflected in the uncertainty in $\hat{\beta}$ ($\hat{\mu}$) and \hat{S}_{uk} (\hat{S}_{ok}). One might more accurately refer to $\hat{\beta}$

($\hat{\mu}$) as the EBLUE and \hat{S}_{uk} (\hat{S}_{ok}) as the EBLUP in those circumstances where estimators of θ and τ^2 are used, where the “E” refers to “empirical”.

3.3 Estimation of $\{\theta, \tau^2\}$

Recall that the parameters in θ govern the covariance function of $S(\cdot)$; that is, $C(\cdot, \cdot) \equiv C(\cdot, \cdot; \theta)$. Therefore, θ also governs the semivariogram $\gamma(\cdot, \cdot)$ of $S(\cdot)$, since

$$\gamma(s, u) = \frac{1}{2}[C(s, s) + C(u, u)] - C(s, u); \quad s, u \in D.$$

That is, we can write $\gamma(\cdot, \cdot) \equiv \gamma(\cdot, \cdot; \theta)$. Equations (2.8) and (2.9) show that the parameters $\{\theta, \tau^2\}$ govern $C_z(\cdot, \cdot)$ and $\gamma_z(\cdot, \cdot)$, the covariance function and semivariogram, respectively, of $Z(\cdot)$.

In practice, the parameters $\{\theta, \tau^2\}$ are estimated by “estimating” the covariance function or the semivariogram. Cressie (1993a, Sect. 2.4.1) makes the case that semivariogram “estimation” is to be preferred.

How is the semivariogram “estimated”? This is done in two steps: First, estimate the semivariogram nonparametrically (e.g., by the method-of-moments estimator) at selected lags $\{h_1, \dots, h_k\}$ (Cressie, 1993a, Sect. 2.4). Second, fit a valid parametric semivariogram model to the nonparametric estimates (Cressie, 1993a, Sect. 2.6). The assumption is that this two-step estimation procedure yields $\gamma_z(\cdot, \cdot; \{\hat{\theta}, \hat{\tau}^2\})$, from which the parameter estimates $\{\hat{\theta}, \hat{\tau}^2\}$ may be taken.

Consider the S -process defined in (2.3). So far, $S(\cdot)$ has been decomposed into components reflecting two scales of variation, that is, large-scale variation (represented by $\mu(\cdot)$) and small-scale variation (represented by $\delta(\cdot)$). Here, following Cressie (1993a, Sect. 3.1), we decompose $\delta(\cdot)$ into two further components as follows,

$$\delta(s) = W(s) + \eta(s); \quad s \in D, \quad (3.17)$$

where $W(\cdot)$ represents a zero-mean, L_2 -continuous (i.e., $E(W(s+h) - W(s))^2 \rightarrow 0$ as $\|h\| \rightarrow 0$), *smooth small-scale-variation process*, such that $\text{var}(W(s)) \equiv \sigma_W^2$; $s \in D$, and $\eta(\cdot)$ represents a

zero-mean process, independent of $W(\cdot)$, whose range exists and is smaller than $\min\{\|s_i - s_j\| : 1 \leq i < j \leq n\}$. Call $\eta(\cdot)$ the *microscale process*, and denote $\text{var}(\eta(s)) \equiv \sigma_\eta^2$; $s \in D$.

The η -process represents the variation in $S(\cdot)$ at scales too small to be measured, and this includes the variation due to possible discontinuities in $S(\cdot)$. Discontinuities in $S(\cdot)$ might occur, for example, if $S(\cdot)$ represents the quantity of gold in an ore-body and the gold is distributed in nuggets; the variation due to these discontinuities is called the “nugget” variance of $S(\cdot)$, and this is defined as

$$\sigma_0^2 \equiv \lim_{\|h\| \rightarrow 0} \gamma(h). \quad (3.18)$$

Note that $\sigma_0^2 \leq \sigma_\eta^2$, although it is often assumed that $\sigma_0^2 = \sigma_\eta^2$ (e.g., Matheron, 1963). If measurement error is present, then we observe $Z(\cdot)$, and the nugget variance of $Z(\cdot)$, which includes the measurement-error variance, is defined as

$$\tau_0^2 \equiv \lim_{\|h\| \rightarrow 0} \gamma_z(h) = \lim_{\|h\| \rightarrow 0} \gamma(h) + \tau^2 = \sigma_0^2 + \tau^2. \quad (3.19)$$

For example, we can now write

$$\theta = (\sigma_0^2, \sigma_\eta^2, \sigma_W^2, \mathbf{a}')',$$

where \mathbf{a} is the *range* of $S(\cdot)$ (i.e., the smallest lag in the direction $\mathbf{a}/\|\mathbf{a}\|$ at which $\text{cov}(S(s), S(s+\mathbf{a})) = 0$).

In practice, there should be no problem estimating the range \mathbf{a} , but it may be difficult to estimate the individual components of variation, τ^2 , σ_0^2 , σ_η^2 , or σ_W^2 . If second-order stationarity holds, then we can estimate the *sill* of $Z(\cdot)$, which is $C_z(0) = \sigma_W^2 + \sigma_\eta^2 + \tau^2$. In addition, extrapolation of the estimated semivariogram at lags closest to zero yields an estimate of $\sigma_\eta^2 + \tau^2$, rather than of $\tau_0^2 \equiv \sigma_0^2 + \tau^2$ (Cressie, 1988). Even if we are willing to assume that $\sigma_0^2 = \sigma_\eta^2$ (i.e., $\tau_0^2 = \sigma_\eta^2 + \tau^2$), the estimated semivariogram can tell us nothing about the nature of the measurement error τ^2 . Therefore, unless we are also willing to assume that $\sigma_0^2 = 0$, we need an independent estimate of τ^2 . If replicate observations are available, then τ^2 may be directly estimated from the “pure error” variance of the replicates. Another possibility is to estimate τ^2 from the measuring instrument itself.

3.4 Interpolation and Measurement Error

If measurement-error model (2.2) is assumed, then the universal (or ordinary) kriging predictor “filters out” the measurement error at prediction locations that coincide with sampling locations.

Proposition 3.1 *Suppose we wish to predict $S(s_i)$, for some $s_i \in A$, where A is the sample of locations upon which $Z(\cdot)$ is observed. Then*

$$\hat{S}_{uk}(s_i) = Z(s_i)$$

if and only if $\tau^2 = 0$.

Proof: Recall that the universal kriging predictor of $S(s_i)$ is

$$\hat{S}_{uk}(s_i) = \mathbf{x}(s_i)' \hat{\beta} + \mathbf{c}(s_i)' \Sigma^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}),$$

where $\mathbf{x}(s_i)$, \mathbf{X} , and $\hat{\beta}$ are defined as in (3.11) and (3.12), $\Sigma = \text{var}(\mathbf{Z})$, and

$$\mathbf{c}(s_i) = (C(s_1, s_i), \dots, C(s_i, s_i), \dots, C(s_n, s_i))'.$$

Define

$$\Sigma_s \equiv \text{var}(S(s_1), \dots, S(s_n));$$

that is, the (i, j) th element of Σ_s is $C(s_i, s_j)$. Then $\mathbf{c}(s_i)$ is the i th column (or row) of Σ_s , and

$$\mathbf{c}(s_i)' \Sigma_s^{-1} = \mathbf{e}_i',$$

where \mathbf{e}_i is an $n \times 1$ vector of $(n - 1)$ zeros and a one in the i th position. Further,

$$\Sigma = \Sigma_s + \tau^2 \mathbf{I}.$$

where \mathbf{I} is an $n \times n$ identity matrix, and $\tau^2 = \text{var}(\epsilon(s))$; $s \in D$.

Suppose for the moment that $\tau^2 = 0$. Then $\Sigma = \Sigma_s$, and

$$\begin{aligned} \hat{S}_{uk}(s_i) &= \mathbf{x}(s_i)' \hat{\beta} + \mathbf{e}_i' (\mathbf{Z} - \mathbf{X} \hat{\beta}) \\ &= \mathbf{x}(s_i)' \hat{\beta} + Z(s_i) - \mathbf{x}(s_i)' \hat{\beta} \\ &= Z(s_i). \end{aligned}$$

Now suppose that $\tau^2 > 0$. Then by the following identity (e.g., Mardia et al. 1979, App. A.2.4),

$$\Sigma^{-1} = \Sigma_s^{-1} - \Sigma_s^{-1}(\Sigma_s^{-1} + \tau^{-2}\mathbf{I})^{-1}\Sigma_s^{-1},$$

we have

$$\mathbf{c}(\mathbf{s}_i)' \Sigma^{-1} = \mathbf{e}_i'(\mathbf{I} - (\Sigma_s^{-1} + \tau^{-2}\mathbf{I})^{-1}\Sigma_s^{-1}).$$

Hence,

$$\begin{aligned} \hat{S}_{uk}(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)' \hat{\beta} + \mathbf{e}_i'(\mathbf{I} - (\Sigma_s^{-1} + \tau^{-2}\mathbf{I})^{-1}\Sigma_s^{-1})(\mathbf{Z} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{Z}(\mathbf{s}_i) - \mathbf{e}_i'(\Sigma_s^{-1} + \tau^{-2}\mathbf{I})^{-1}\Sigma_s^{-1}(\mathbf{Z} - \mathbf{X}\hat{\beta}) \\ &\neq \mathbf{Z}(\mathbf{s}_i), \end{aligned}$$

since $\tau^2 > 0$. Therefore, universal kriging is an *exact interpolator* if and only if $\tau^2 = 0$. \square

It is not difficult to show that Proposition 3.1 applies similarly to ordinary kriging.

3.5 Kriging and Uncorrelated Spatial Processes

Consider the measurement-error spatial model (2.2), with $S(\cdot)$ and $\epsilon(\cdot)$ defined exactly as in Section 2, but with

$$\text{cov}[\delta(\mathbf{s}), \delta(\mathbf{u})] = \begin{cases} C(\mathbf{s}, \mathbf{s}); & \mathbf{s} = \mathbf{u}, \\ 0; & \mathbf{s} \neq \mathbf{u}. \end{cases}$$

Thus,

$$C_z(\mathbf{s}, \mathbf{u}) = \begin{cases} C(\mathbf{s}, \mathbf{s}) + \tau^2; & \mathbf{s} = \mathbf{u}, \\ 0; & \mathbf{s} \neq \mathbf{u}, \end{cases}$$

where $C_z(\mathbf{s}, \mathbf{u}) = \text{cov}[Z(\mathbf{s}), Z(\mathbf{u})]$. That is, $S(\cdot)$ (and hence $Z(\cdot)$) is a spatially uncorrelated process.

For such an $S(\cdot)$ and $B \subset D$, satisfying $|B| > 0$,

$$\begin{aligned} \mathbf{c}(B) &\equiv (C(\mathbf{s}_1, B), \dots, C(\mathbf{s}_n, B))' \\ &= \mathbf{0}. \end{aligned}$$

This is because

$$\begin{aligned} C(s_i, B) &= \frac{1}{|B|} \int_B C(s_i, u) du; \quad i = 1, \dots, n \\ &= 0; \quad i = 1, \dots, n, \end{aligned} \quad (3.20)$$

since $C(s_i, u)$ in (3.20) is positive for at most one point location in B ; $i = 1, \dots, n$.

In practice, B is discretized into a grid of N locations, yielding $B = \{u_1, \dots, u_N\}$. Then

$$C(s_i, B) = \frac{1}{N} \sum_{j=1}^N C(s_i, u_j); \quad i = 1, \dots, n,$$

where $C(s_i, u_j) = 0$, if $s_i \neq u_j$; $i = 1, \dots, n$, $j = 1, \dots, N$.

Proposition 3.2 *Suppose that $S(\cdot)$ is a spatially uncorrelated process, $B = \{u_1, \dots, u_N\} \subset D$, and denote $A^* \equiv B \cap A$, where $A = \{s_1, \dots, s_n\}$. Then the universal kriging predictor of $S(B)$ is*

$$\hat{S}_{uk}(B) = \mathbf{x}(B)' \hat{\beta} + \frac{1}{N} \sum_{i=1}^n C(s_i, s_i) I(s_i \in A^*) \mathbf{e}_i' \Sigma^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}), \quad (3.21)$$

and the ordinary kriging predictor of $S(B)$ is

$$\hat{S}_{ok}(B) = \hat{\mu} + \frac{1}{N} \sum_{i=1}^n C(s_i, s_i) I(s_i \in A^*) \mathbf{e}_i' \Sigma^{-1} (\mathbf{Z} - \hat{\mu} \mathbf{1}), \quad (3.22)$$

where \mathbf{e}_i is an $n \times 1$ vector of $(n-1)$ zeros and a one in the i th position, and $I(\cdot)$ is the indicator function.

Proof: Since $A^* \equiv B \cap A$,

$$\mathbf{c}(B) = \frac{1}{N} \sum_{i=1}^n C(s_i, s_i) I(s_i \in A^*) \mathbf{e}_i, \quad (3.23)$$

yielding the universal and ordinary kriging predictors (3.21) and (3.22), respectively. \square

Several consequences of Proposition 3.2 are noted here:

1. If $s_i \notin B$; $i = 1, \dots, n$, then $s_i \notin A^*$; $i = 1, \dots, n$. Thus $I(s_i \in A^*) = 0$; $i = 1, \dots, n$, hence the universal kriging predictor of $S(B)$ is

$$\hat{S}_{uk}(B) = \mathbf{x}(B)' \hat{\beta},$$

and the ordinary kriging predictor of $S(B)$ is

$$\hat{S}_{ok}(B) = \hat{\mu}.$$

2. If $A^* = \{s_1\}$, then

$$\begin{aligned}\hat{S}_{uk}(s_1) &= \mathbf{x}(s_1)' \hat{\beta} + C(s_1, s_1) \mathbf{e}_1' \Sigma^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}), \\ &= \mathbf{x}(s_1)' \hat{\beta} + (C(s_1, s_1) / (C(s_1, s_1) + \tau^2)) \mathbf{e}_1' (\mathbf{Z} - \mathbf{X} \hat{\beta}), \\ &= \mathbf{x}(s_1)' \hat{\beta} + (C(s_1, s_1) / (C(s_1, s_1) + \tau^2)) (Z(s_1) - \mathbf{x}(s_1)' \hat{\beta}),\end{aligned}$$

since

$$\Sigma_{ij} = \begin{cases} C(s_i, s_i) + \tau^2; & i = j, \\ 0; & i \neq j, \end{cases} \quad (3.24)$$

where Σ_{ij} denotes the (i, j) th element of $\Sigma = \text{var}(\mathbf{Z})$. Similarly,

$$\hat{S}_{ok}(s_1) = \hat{\mu} + (C(s_1, s_1) / (C(s_1, s_1) + \tau^2)) (Z(s_1) - \hat{\mu}).$$

If, in addition, $\tau^2 = 0$, then

$$\hat{S}_{uk}(s_1) = \hat{S}_{ok}(s_1) = Z(s_1),$$

thereby confirming Proposition 3.1.

3. If $A^* = \{s_1\}$, and $C(s, s) = \text{var}(S(s)) \equiv \sigma^2$, independent of s , then

$$\hat{S}_{ok}(s_1) = \frac{\tau^2}{\tau^2 + \sigma^2} \bar{Z} + \frac{\sigma^2}{\tau^2 + \sigma^2} Z(s_1), \quad (3.25)$$

where $\bar{Z} = \mathbf{1}' \mathbf{Z} / n$ is the BLUE of μ . We note in passing the similarity between (3.25) and the posterior expectation of a conjugate normal Bayesian model. For example, suppose $X | \theta \sim N(\theta, \tau^2)$ (i.e., X , conditional on θ , is distributed as a $N(\theta, \tau^2)$ random variable), and $\theta \sim N(\mu, \sigma^2)$. Then

$$E(\theta | X) = \frac{\tau^2}{\tau^2 + \sigma^2} \mu + \frac{\sigma^2}{\tau^2 + \sigma^2} X.$$

3.6 Multivariate Prediction

So far we have considered only univariate spatial processes, that is, where at most one observation is taken at each location in D . We now extend this to the multivariate case, where multiple observations may be taken at a single location.

Consider the vector-valued, m -variate spatial process

$$\mathbf{Z}(\cdot) \equiv \{\mathbf{Z}(s) : s \in D\}, \quad (3.26)$$

where $\mathbf{Z}(s) = (Z_1(s), \dots, Z_m(s))'$; $s \in D$. Define the multivariate measurement-error spatial model

$$Z_{ij}(s) = S_i(s) + \epsilon_{ij}(s); \quad s \in D, \quad j = 1, \dots, m_i, \quad i = 1, \dots, M, \quad (3.27)$$

where M is the number of distinct physical (i.e., signal) processes, represented by $\{S_1(\cdot), \dots, S_M(\cdot)\}$, and $\sum_{i=1}^M m_i = m$.

If $m_1 = \dots = m_M = 1$, then the $M = m$ distinct signal processes $\{S_1(\cdot), \dots, S_M(\cdot)\}$ are of interest, and observations are taken from the corresponding measurement-error contaminated processes, $\{Z_{11}(\cdot), \dots, Z_{M1}(\cdot)\}$, (e.g., for $m = M = 2$, suppose that lead and zinc, each with its own source of measurement-error, are observed in an ore body). If $m_i > 1$, for some $i = 1, \dots, M$, then the i th physical process is measured by $m_i > 1$ different measurement methods (e.g., for $M = 1$ and $m_1 = 2$, suppose that ground temperature is measured by ground-based measuring instruments, as well as by satellite infrared imagery).

The signal processes can be further decomposed as

$$S_i(s) = \mu_i(s) + \delta_i(s); \quad s \in D, \quad i = 1, \dots, M, \quad (3.28)$$

where $\mu_i(\cdot)$ is the large-scale, deterministic mean structure, and $\delta_i(\cdot)$ is the small-scale, stochastic structure of the i th signal process; $i = 1, \dots, M$. If each $\mu_i(\cdot)$ is linear in unknown parameters β_i ; $i = 1, \dots, M$, then

$$\mu_i(s) = \sum_{j=1}^{p_i+1} x_{i,j-1}(s) \beta_{i,j-1}, \quad (3.29)$$

where $\beta_i = (\beta_{i0}, \dots, \beta_{ip_i})'$ is the i th vector of unknown parameters, and $\{x_{i0}(\cdot), \dots, x_{ip_i}(\cdot)\}$ are known functions; $i = 1, \dots, M$. Upon integrating both sides, equation (3.29) can be written as

$$\mu_i(B) = \mathbf{x}_i(B)' \beta_i; \quad B \subset D, \quad (3.30)$$

where $\mathbf{x}_i(B) \equiv (x_{i0}(B), \dots, x_{ip_i}(B))'$, $x_{ij}(B) \equiv \frac{1}{|B|} \int_B x_{ij}(u) du$; $j = 1, \dots, p_i$, $i = 1, \dots, M$. If $B = \{s_0\}$, then (3.29) with $s = s_0$ becomes (3.30) with $\mathbf{x}_i(s_0) \equiv (x_{i0}(s_0), \dots, x_{ip_i}(s_0))'$; $i = 1, \dots, M$.

Assume that for each $\delta_i(\cdot)$; $i = 1, \dots, M$,

$$E[\delta_i(s)] = 0; \quad s \in D, \quad i = 1, \dots, M \quad (3.31)$$

$$\text{cov}[\delta_i(s), \delta_j(u)] \equiv C_{ij}(s, u); \quad s, u \in D, \quad i, j = 1, \dots, M, \quad (3.32)$$

where $C_{ij}(\cdot, \cdot)$ is the *cross-covariance function* of $(\delta_i(\cdot), \delta_j(\cdot))$ (i.e., of $(S_i(\cdot), S_j(\cdot))$), for all $i \neq j$, and $C_{ii}(\cdot, \cdot)$ is the *covariance function* of $\delta_i(\cdot)$ (i.e., of $S_i(\cdot)$), $i = 1, \dots, M$. Note that it is possible that $C_{ij}(s, u) \neq C_{ji}(s, u)$ (e.g., Journel and Huijbregts, 1978, p. 41).

In addition, for each $j = 1, \dots, m_i$, $i = 1, \dots, M$, $\epsilon_{ij}(\cdot)$ is an independent zero-mean, white-noise process, with

$$\text{var}(\epsilon_{ij}(s)) \equiv \tau_{ij}^2; \quad s \in D, \quad j = 1, \dots, m_i, \quad i = 1, \dots, M, \quad (3.33)$$

and $S_i(\cdot)$ and $\epsilon_{ij}(\cdot)$ are assumed to be independent; $j = 1, \dots, m_i$, $i = 1, \dots, M$.

3.6.1 Cokriging

Assume for simplicity that $m_1 = \dots = m_M = 1$ and hence $M = m$. Then we can write the multivariate measurement-error spatial model as,

$$Z_i(s) = S_i(s) + \epsilon_i(s); \quad s \in D, \quad i = 1, \dots, m. \quad (3.34)$$

Suppose that the data consist of observations,

$$\mathbf{Z} = (\mathbf{Z}'_1, \dots, \mathbf{Z}'_m)',$$

where $\mathbf{Z}_i = (Z_i(s_{i1}), \dots, Z_i(s_{in_i}))'$; $i = 1, \dots, m$.

Then we can write,

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_m \end{bmatrix},$$

$\mathbf{X}_i = (\mathbf{x}_i(\mathbf{s}_{i1}), \dots, \mathbf{x}_i(\mathbf{s}_{in_i}))'$, $\mathbf{x}_i(\mathbf{s}) \equiv (x_{i0}(\mathbf{s}), \dots, x_{ip_i}(\mathbf{s}))'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_m)'$, $\boldsymbol{\beta}_i = (\beta_{i0}, \dots, \beta_{ip_i})'$, $\boldsymbol{\delta} = (\boldsymbol{\delta}'_1, \dots, \boldsymbol{\delta}'_m)'$, $\boldsymbol{\delta}_i = (\delta_i(\mathbf{s}_{i1}), \dots, \delta_i(\mathbf{s}_{in_i}))'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_1, \dots, \boldsymbol{\epsilon}'_m)'$, $\boldsymbol{\epsilon}_i = (\epsilon_i(\mathbf{s}_{i1}), \dots, \epsilon_i(\mathbf{s}_{in_i}))'$; $i = 1, \dots, m$.

Suppose we wish to predict

$$\mathbf{S} \equiv (\mathbf{S}_1(B), \dots, \mathbf{S}_m(B))'.$$

Ver Hoef and Cressie (1993) define the multivariate BLUP, sometimes referred to as the *cokriging predictor* $\hat{\mathbf{S}}_{co}$, of \mathbf{S} as follows: Suppose $\hat{\mathbf{S}}_{co}$ is unbiased for \mathbf{S} and linear in \mathbf{Z} . Then $\hat{\mathbf{S}}_{co}$ is the cokriging predictor of \mathbf{S} if

$$\text{MSPE}(\hat{\mathbf{S}}) - \text{MSPE}(\hat{\mathbf{S}}_{co}) = E[(\hat{\mathbf{S}} - \mathbf{S})(\hat{\mathbf{S}} - \mathbf{S})'] - E[(\hat{\mathbf{S}}_{co} - \mathbf{S})(\hat{\mathbf{S}}_{co} - \mathbf{S})']$$

is non-negative definite (n.n.d.) for every predictor $\hat{\mathbf{S}}$ of \mathbf{S} that is unbiased and linear in \mathbf{Z} . They show that the cokriging predictor of \mathbf{S} can be written as

$$\hat{\mathbf{S}}_{co} \equiv \mathbf{X}(B)' \hat{\boldsymbol{\beta}} + \mathbf{C}(B)' \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X} \hat{\boldsymbol{\beta}}), \quad (3.35)$$

where

$$\mathbf{X}(B)' = \begin{bmatrix} \mathbf{x}_1(B)' & \mathbf{0}' & \dots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{x}_2(B)' & \dots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \dots & \mathbf{x}_m(B)' \end{bmatrix},$$

$\mathbf{x}_i(B)$ is defined as in (3.30), $\hat{\beta} \equiv (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Z}$,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \dots & \Sigma_{mm} \end{bmatrix},$$

$\Sigma_{ij} = \text{cov}(\delta_i, \delta_j)$, an $n_i \times n_j$ matrix whose (k, l) th element is $C_{ij}(s_{ik}, s_{jl})$; $i \neq j$, $\Sigma_{ii} = \text{var}(\delta_i) + \text{var}(\epsilon_i)$, an $n_i \times n_i$ matrix whose (k, l) th element is $C_{ii}(s_{ik}, s_{il}) + \tau_i^2 I(k = l)$; $i = 1, \dots, m$,

$$\begin{aligned} \mathbf{C}(B) &= \text{cov}(\mathbf{Z}, \mathbf{S}) \\ &= \begin{bmatrix} \mathbf{c}_{11}(B) & \mathbf{c}_{12}(B) & \dots & \mathbf{c}_{1m}(B) \\ \mathbf{c}_{21}(B) & \mathbf{c}_{22}(B) & \dots & \mathbf{c}_{2m}(B) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{c}_{m1}(B) & \mathbf{c}_{m2}(B) & \dots & \mathbf{c}_{mm}(B) \end{bmatrix}, \end{aligned}$$

$\mathbf{c}_{ij}(B) = \text{cov}(\mathbf{Z}_i, \mathbf{S}_j(B)) = (C_{ij}(s_{i1}, B), \dots, C_{ij}(s_{in_i}, B))'$, $C_{ij}(s, B) = \frac{1}{|B|} \int_B C_{ij}(s, u) du$; $i, j = 1, \dots, m$.

Notice that the off-diagonal block matrices of Σ (i.e., Σ_{ij} ; $i \neq j$) consist of cross-covariance functions and the diagonal block matrices of Σ (i.e., Σ_{ii} ; $i = 1, \dots, m$) consist of covariance functions, plus the appropriate measurement-error variance τ_i^2 present in the diagonal elements of Σ_{ii} ; $i = 1, \dots, m$. The matrix $\mathbf{C}(B)$ is not a function of $\{\tau_i^2 : i = 1, \dots, m\}$; the elements of $\mathbf{c}_{ij}(B)$ are obtained from cross-covariance functions; $i \neq j$, and the elements of $\mathbf{c}_{ii}(B)$ are obtained from covariance functions; $i = 1, \dots, m$.

Ver Hoef and Cressie (1993) show that the MSPE matrix of $\hat{\mathbf{S}}_{co}$ is

$$\begin{aligned} \text{MSPE}(\hat{\mathbf{S}}_{co}) &= \Sigma_0 - \mathbf{C}(B)' \Sigma^{-1} \mathbf{C}(B) \\ &\quad + (\mathbf{X}(B)' - \mathbf{C}(B)' \Sigma^{-1} \mathbf{X})(\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{X}(B) - \mathbf{X}' \Sigma^{-1} \mathbf{C}(B)), \end{aligned}$$

where $\Sigma_0 \equiv \text{var}(\mathbf{S})$. It can also be shown (e.g., Ver Hoef and Cressie, 1993) that for any $m \times 1$ vector \mathbf{a} , the cokriging predictor of $\mathbf{a}'\mathbf{S}$ is $\mathbf{a}'\hat{\mathbf{S}}_{co}$. For example, the cokriging predictor of $S_1(B)$ is $\mathbf{e}_1' \hat{\mathbf{S}}_{co}$, where $\mathbf{e}_1 = (1, 0, \dots, 0)'$. That is,

$$\hat{S}_{1,co}(B) \equiv \mathbf{e}_1' \hat{\mathbf{S}}_{co} = \mathbf{x}_1(B)' \hat{\beta}_1 + \mathbf{c}_1(B)' \Sigma^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}),$$

where $\hat{\beta}_1 = (\mathbf{X}_1' \Sigma_{11}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1' \Sigma_{11}^{-1} \mathbf{Z}_1$, and $\mathbf{c}_1(B) = (\mathbf{c}_{11}(B)', \dots, \mathbf{c}_{m1}(B)')'$. This means that the information in the spatial correlation that may exist among *all* of the signal processes $\{S_1(\cdot), \dots, S_m(\cdot)\}$ is used in the cokriging predictor $\hat{S}_{1,co}(B)$. If no spatial correlation exists between $S_1(\cdot)$ and $\{S_2(\cdot), \dots, S_m(\cdot)\}$, then $\hat{S}_{1,co}$ is equal to the universal kriging predictor of $S_1(B)$,

$$\hat{S}_{1,uk}(B) = \mathbf{x}_1(B)' \hat{\beta}_1 + \mathbf{c}_{11}(B)' \Sigma_{11}^{-1} (\mathbf{Z}_1 - \mathbf{X}_1 \hat{\beta}_1),$$

since $\mathbf{c}_{i1}(B) = 0$; $i \neq 1$, and all the elements of the block-matrices Σ_{i1} and Σ_{1i} are zero; $i \neq 1$.

3.6.2 Generalized Universal Kriging

Suppose that $M = 1$, and $m_1 \equiv m > 1$; that is, the multivariate measurement-error spatial model is

$$Z_i(\mathbf{s}) = S(\mathbf{s}) + \epsilon_i(\mathbf{s}); \quad \mathbf{s} \in D, \quad i = 1, \dots, m, \quad (3.36)$$

where $m > 1$,

$$S(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}); \quad \mathbf{s} \in D, \quad (3.37)$$

and, following Section 3.2.2,

$$\mu(B) = \mathbf{x}(B)' \beta; \quad B \subset D, \quad (3.38)$$

with the terms in (3.38) defined as in (3.11). Assume that $\delta(\cdot)$ has first two moments,

$$E[\delta(\mathbf{s})] = 0; \quad \mathbf{s} \in D, \quad (3.39)$$

$$\text{cov}[\delta(\mathbf{s}), \delta(\mathbf{u})] \equiv C(\mathbf{s}, \mathbf{u}); \quad \mathbf{s}, \mathbf{u} \in D. \quad (3.40)$$

Suppose that the data consist of observations

$$\mathbf{Z} = (\mathbf{Z}'_1, \dots, \mathbf{Z}'_m)',$$

where $\mathbf{Z}_i = (Z_i(\mathbf{s}_{i1}), \dots, Z_i(\mathbf{s}_{in_i}))'$; $i = 1, \dots, m$. Then we can write

$$\mathbf{Z} = \mathbf{X}\beta + \delta + \epsilon,$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix},$$

$\mathbf{X}_i = (\mathbf{x}(s_{i1}), \dots, \mathbf{x}(s_{in_i}))'$; $i = 1, \dots, m$, $\mathbf{x}(s) \equiv (x_0(s), \dots, x_p(s))'$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$, $\boldsymbol{\delta} = (\delta'_1, \dots, \delta'_m)'$, $\boldsymbol{\delta}_i = (\delta(s_{i1}), \dots, \delta(s_{in_i}))'$; $i = 1, \dots, m$, $\boldsymbol{\epsilon} = (\epsilon'_1, \dots, \epsilon'_m)'$, $\epsilon_i = (\epsilon(s_{i1}), \dots, \epsilon(s_{in_i}))'$; $i = 1, \dots, m$. Suppose we wish to predict $S(B)$. It is not difficult to show that the BLUP, or for want of a better term, the *generalized universal kriging* predictor, of $S(B)$ is

$$\hat{S}_{gu}(B) \equiv \mathbf{x}(B)' \hat{\boldsymbol{\beta}} + \mathbf{c}(B)' \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X} \hat{\boldsymbol{\beta}}), \quad (3.41)$$

where $\mathbf{x}(B)$ is defined as in (3.11), $\hat{\boldsymbol{\beta}} \equiv (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}$,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \dots & \boldsymbol{\Sigma}_{1m} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \dots & \boldsymbol{\Sigma}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{m1} & \boldsymbol{\Sigma}_{m2} & \dots & \boldsymbol{\Sigma}_{mm} \end{bmatrix},$$

$\boldsymbol{\Sigma}_{ij} = \text{cov}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j)$, an $n_i \times n_j$ matrix whose (k, l) th element is $C(s_{ik}, s_{jl})$; $i \neq j$, $\boldsymbol{\Sigma}_{ii} = \text{var}(\boldsymbol{\delta}_i) + \tau_i^2 \mathbf{I}$; $i = 1, \dots, m$, $\mathbf{c}(B) = \text{cov}(\mathbf{Z}, S(B)) = (C(B, s_{11}), \dots, C(B, s_{1n_1}), C(B, s_{21}), \dots, C(B, s_{2n_2}), \dots, C(B, s_{m1}), \dots, C(B, s_{mn_m}))'$.

It is easy to show that the MSPE of $\hat{S}_{gu}(B)$ is

$$\begin{aligned} \text{MSPE}(\hat{S}_{gu}(B)) &= C(B, B) - \mathbf{c}(B)' \boldsymbol{\Sigma}^{-1} \mathbf{c}(B) \\ &\quad + (\mathbf{x}(B)' - \mathbf{c}(B)' \boldsymbol{\Sigma}^{-1} \mathbf{X}) (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} (\mathbf{x}(B) - \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{c}(B)), \end{aligned}$$

where $C(B, B) = \text{var}(S(B))$.

Observe that $\hat{S}_{gu}(B)$ and $\text{MSPE}(\hat{S}_{gu}(B))$ contain *no cross-covariance functions*, and these quantities differ from the universal kriging predictor (3.13) and its MSPE (3.14) only in the m *distinct* measurement-error variances $\{\tau_i^2 : i = 1, \dots, m\}$ added to the diagonal elements of $\boldsymbol{\Sigma}$, in accordance with the model (3.36). Hence we have used the term “generalized” universal

kriging; Carroll (1995) provides an example where generalized simple kriging was used in accordance with model (3.36).

3.6.3 Generalized Cokriging

Suppose $M > 1$ and not all of $\{m_i : i = 1, \dots, M\}$ are equal to one. Suppose we want to predict

$$\mathbf{S} \equiv (S_1(B), \dots, S_M(B))'.$$

For want of a better term, we could refer to the multivariate BLUP of \mathbf{S} as the *generalized cokriging predictor*. This would have the form of (3.35), except that in submatrices corresponding to the cases where signal processes are measured by more than one method, the features of (3.41) will apply. This means that care must be taken to specify the multivariate measurement-error spatial model correctly, so that the matrix Σ can be constructed correctly and the correct distinction between cross-covariance functions and covariance functions can be made.

3.7 Nonlinear Predictors

Suppose that $(S(\cdot), \epsilon(\cdot))$ is not a bivariate Gaussian process. Then

$$\hat{S}_{sk}(B) \neq \hat{S}_{bp}(B)$$

and, depending on the degree of non Gaussianity, $\hat{S}_{sk}(B)$ (likewise $\hat{S}_{uk}(B)$) may be far from optimal. Such cases have motivated the search for alternative predictors with better optimality properties than $\hat{S}_{sk}(B)$ and $\hat{S}_{uk}(B)$, and three such predictors, nonlinear in \mathbf{Z} , are presented in this section.

In general, the best predictor of $S(B)$, $\hat{S}_{bp}(B)$, is *nonlinear* in \mathbf{Z} , and can be written as

$$\hat{S}_{bp}(B) \equiv E[S(B)|\mathbf{Z}] = \int t dG_S(t|\mathbf{Z}), \quad (3.42)$$

where

$$\begin{aligned} G_S(t|\mathbf{Z}) &\equiv \Pr(S(B) \leq t|\mathbf{Z}); \quad t \in \mathbb{R} \\ &= E[I(S(B) \leq t|\mathbf{Z})]; \quad t \in \mathbb{R}. \end{aligned}$$

Notice that $G_S(t|\mathbf{Z})$ is the *best predictor* of $I(S(B) \leq t)$.

3.7.1 Indicator Kriging

The idea behind indicator kriging is to approximate \hat{S}_{bp} by “estimating” $G_S(t|\mathbf{Z})$. In his original formulation of indicator kriging, Journel (1983) considered only the case $B = \{\mathbf{s}_0\}$, and he considered the predictand $Z(\mathbf{s}_0)$. (In practice, this is equivalent to assuming that no measurement error exists, that is, $Z(\cdot) \equiv S(\cdot)$). In such a case, the *indicator kriging predictor* of $Z(\mathbf{s}_0)$ is defined as

$$\hat{Z}_{ik}(\mathbf{s}_0) \equiv \int t d\hat{G}_{ik}(t|\mathbf{Z}), \quad (3.43)$$

where $\hat{G}_{ik}(t|\mathbf{Z})$ is the BLUP (i.e., universal or ordinary kriging predictor) of $I(Z(\mathbf{s}_0) \leq t)$. That is,

$$\hat{G}_{ik}(t|\mathbf{Z}) \equiv \sum_{i=1}^n \lambda_i(t) I(Z(\mathbf{s}_i) \leq t), \quad (3.44)$$

where $\lambda_1(t), \dots, \lambda_n(t)$ are chosen to minimize

$$E \left[\left(\sum_{i=1}^n \lambda_i(t) I(Z(\mathbf{s}_i) \leq t) - I(Z(\mathbf{s}_0) \leq t) \right)^2 \right],$$

with respect to $\lambda_1(t), \dots, \lambda_n(t)$, subject to $E[\sum_{i=1}^n \lambda_i(t) I(Z(\mathbf{s}_i) \leq t)] = E[I(Z(\mathbf{s}_0) \leq t)]$, for each $t \in \mathbb{R}$. Notice that $\lambda_1(t), \dots, \lambda_n(t)$ must be calculated anew for *each* t under consideration.

Journel (1983) points out that $\hat{G}_{ik}(t|\mathbf{Z})$ is the BLUP of $I(Z(\mathbf{s}_0) \leq t)$, and that $G_Z(t|\mathbf{Z}) \equiv \Pr(Z(\mathbf{s}_0) \leq t|\mathbf{Z})$; $t \in \mathbb{R}$, is the best predictor of $I(Z(\mathbf{s}_0) \leq t)$, and this provides his justification for using $\hat{G}_{ik}(t|\mathbf{Z})$ as an “estimator” of $G_Z(t|\mathbf{Z})$. But there are problems with this approach, and these are listed here.

Problems with Indicator Kriging

1. Optimality of $\hat{G}_{ik}(t|\mathbf{Z})$: We have that $\hat{G}_{ik}(t|\mathbf{Z})$ is the BLUP of $I(Z(\mathbf{s}_0) \leq t)$, and that $G_Z(t|\mathbf{Z})$ is the best predictor of $I(Z(\mathbf{s}_0) \leq t)$. Therefore,

$$E[\hat{G}_{ik}(t|\mathbf{Z})] = E[G_Z(t|\mathbf{Z})];$$

that is, $\hat{G}_{ik}(t|\mathbf{Z})$ is an unbiased “estimator” of $G_Z(t|\mathbf{Z})$. But, $\{I(Z(\mathbf{s}_i) \leq t) : i = 1, \dots, n\}$ are Bernoulli random variables, which are very different from Gaussian random variables. Thus, $\hat{G}_{ik}(t|\mathbf{Z})$ may be far from optimal as a predictor of $I(Z(\mathbf{s}_0) \leq t)$ (Cressie, 1993b).

2. Change of support: If B is not a point location, then it is not a straightforward problem to obtain the BLUP of $I(Z(B) \leq t)$ from data $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$ on point support, since

$$\begin{aligned} I(Z(B) \leq t) &= I\left(\frac{1}{|B|} \int_B Z(\mathbf{u}) d\mathbf{u} \leq t\right) \in \{0, 1\} \\ &\neq \frac{1}{|B|} \int_B I(Z(\mathbf{u}) \leq t) d\mathbf{u} \in [0, 1]. \end{aligned}$$

Thus, additional assumptions have to be made about $\text{cov}((I(Z(\mathbf{s}_1) \leq t), \dots, I(Z(\mathbf{s}_n) \leq t))', I(Z(B) \leq t))$, since it is not enough to specify just $\text{cov}(I(Z(\mathbf{s}_i) \leq t), I(Z(\mathbf{s}_0) \leq t)) : \mathbf{s}_0 \in B$ (Journel, 1983).

3. Measurement error and bias: Assume measurement-error model (2.2) with $\tau^2 > 0$. Then,

$$\begin{aligned} E[\hat{G}_{ik}(t|\mathbf{Z})] &= E[I(Z(\mathbf{s}_0) \leq t)] = \Pr(Z(\mathbf{s}_0) \leq t) \\ &= \Pr(S(\mathbf{s}_0) + \epsilon(\mathbf{s}_0) \leq t) \\ &= G_S * G_\epsilon(t) \\ &\neq G_S(t) = E[I(S(\mathbf{s}_0) \leq t)] \\ &= E[G_S(t|\mathbf{Z})], \end{aligned}$$

where $G_S(t) = \Pr(S(\mathbf{s}_0) \leq t)$, $G_\epsilon(t) = \Pr(\epsilon(\mathbf{s}_0) \leq t)$, and “*” denotes convolution. That is, $E[\hat{G}_{ik}(t|\mathbf{Z})] \neq E[G_S(t|\mathbf{Z})]$, or $\hat{G}_{ik}(t|\mathbf{Z})$ is a biased “estimator” of $G_S(t|\mathbf{Z})$. Consequently, $\hat{Z}_{ik}(\mathbf{s}_0)$ is a biased predictor of $S(\mathbf{s}_0)$, and the bias may be substantial if τ^2 is large.

3.7.2 Indicator Cokriging

Lajaunie (1990) notes that the information in $\{I(Z(\mathbf{s}_i) \leq t) : i = 1, \dots, n\}$ is vastly inferior to the information in $\{Z(\mathbf{s}_i) : i = 1, \dots, n\}$, but that the indicator information can be enriched

considerably by adding extra thresholds to obtain

$$\{I(Z(\mathbf{s}_i) \leq t_j) : j = 1, \dots, m_i; i = 1, \dots, n\}. \quad (3.45)$$

Define

$$\hat{G}_{ic}(t|\mathbf{Z}) \equiv \sum_{i=1}^n \sum_{j=1}^{m_i} \lambda_i(t_j) I(Z(\mathbf{s}_i) \leq t_j), \quad (3.46)$$

where $\{\lambda_i(t_j) : j = 1, \dots, m_i; i = 1, \dots, n\}$ are chosen to minimize

$$E \left[\left(\sum_{i=1}^n \sum_{j=1}^{m_i} \lambda_i(t_j) I(Z(\mathbf{s}_i) \leq t_j) - I(Z(\mathbf{s}_0) \leq t_j) \right)^2 \right],$$

with respect to $\{\lambda_i(t_j) : j = 1, \dots, m_i; i = 1, \dots, n\}$, subject to $E[\sum_{i=1}^n \sum_{j=1}^{m_i} \lambda_i(t_j) I(Z(\mathbf{s}_i) \leq t_j)] = E[I(Z(\mathbf{s}_0) \leq t_j)]$, for each $t_j \in \mathbb{R}; j = 1, \dots, m_i, i = 1, \dots, n$.

Then $\hat{G}_{ic}(t|\mathbf{Z})$ is the cokriging predictor (see Section 3.6.1) of $I(Z(\mathbf{s}_0) \leq t)$, and it is *always* superior to $\hat{G}_{ik}(t|\mathbf{Z})$, since the σ -algebra generated by $\{I(Z(\mathbf{s}_i) \leq t) : i = 1, \dots, n\}$ is contained in the σ -algebra generated by (3.45) (Cressie, 1993b). This yields an “improved” predictor of $Z(\mathbf{s}_0)$, called the *indicator cokriging predictor*, which is defined as

$$\hat{Z}_{ic}(\mathbf{s}_0) \equiv \int t d\hat{G}_{ic}(t|\mathbf{Z}), \quad (3.47)$$

where $\hat{G}_{ic}(t|\mathbf{Z})$ is given by (3.46). Lajaunie (1990) suggests that indicator cokriging performs better than indicator kriging, but the same problems with regard to change of support and measurement error and bias, that exist in the indicator kriging predictor, exist here too.

3.7.3 Disjunctive Kriging

The best predictor $\hat{S}_{bp}(\mathbf{s}_0)$ requires knowledge of the joint distribution of $(S(\mathbf{s}_0), \mathbf{Z}')$ and, in the non-Gaussian case, this is not a trivial problem. Assuming a measurement-error-free model (i.e., $Z(\cdot) \equiv S(\cdot)$), Matheron (1976) pioneered predictors of $Z(\mathbf{s}_0)$ of the form,

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^n f_i(Z(\mathbf{s}_i)), \quad (3.48)$$

where $\{f_i(\cdot) : i = 1, \dots, n\}$ are measurable, square-integrable functions and he called the optimal predictor in this class the *disjunctive kriging predictor*. We denote this predictor as,

$$\hat{Z}_{dk}(\mathbf{s}_0) = \sum_{i=1}^n g_i(Z(\mathbf{s}_i)), \quad (3.49)$$

where $\{g_i(\cdot) : i = 1, \dots, n\}$ are chosen to minimize the mean-squared prediction error of predictors of the form given by (3.48). Matheron (1976) shows that (under certain conditions) only knowledge of the bivariate distributions of $\{(Z(s_i), Z(s_j)) : 0 \leq i < j \leq n\}$ are required.

Now, any measurable function of $Z(s_i)$ can be approximated by simple functions, which are linear combinations of indicator functions (e.g., Loeve, 1977, Section 5.3). Thus, the indicator cokriging predictor given by (3.46) is optimal (up to discretization) in the class of predictors given by (3.48); that is, $\hat{Z}_{ic}(s_0)$ is a discretized version of $\hat{Z}_{dk}(s_0)$.

If $Z(\cdot)$ can only take a finite number of values $\{z_1, \dots, z_k\}$, then

$$\hat{Z}_{dk}(s_0) = \sum_{i=1}^n \sum_{j=1}^k \lambda_i(z_j) I(Z(s_i) \leq z_j), \quad (3.50)$$

in which case, disjunctive kriging and indicator cokriging are exactly equivalent. Clearly then, the problems that we saw with indicator cokriging also apply to disjunctive kriging.

4 Prediction of Nonlinear Functionals of the S -process

In this section, we shall consider predictands of the form

$$h(S(\cdot)) = g(S(B)), \quad (4.1)$$

where g is some nonlinear measurable function, and $B \subset D$. Examples of such predictands might include

$$g(S(B)) = \log(S(B)),$$

or

$$g(S(B)) = I(S(B) \leq t),$$

where $I(\cdot)$ is the indicator function, and t is some specified threshold.

Unbiasedness

When optimality is measured with respect to the squared-error loss function, then a predictor with small risk is one with a small mean-squared prediction error (MSPE). Suppose \hat{S} is a predictor of S . Then

$$\text{MSPE}(\hat{S}) = \text{var}(\hat{S} - S) + [\text{bias}(\hat{S} - S)]^2,$$

where $\text{bias}(\hat{S} - S) \equiv E[\hat{S} - S]$.

What is the attraction of an unbiased predictor with a large MSPE, if there exists a biased predictor with a smaller MSPE? The usefulness of the unbiasedness property becomes apparent when *many* predictions are required (e.g., a “map” requiring predictions at many locations). To see why this is so, consider some summary statistic of those many predictors, such as the average of all of them. The variance of the average of many predictors is never more (usually less) than the variance of any of the individual predictors constituting the average (see Proposition 4.2). This means that by averaging many unbiased predictors to form a summary statistic, the MSPE may be reduced, sometimes substantially. On the other hand, for biased predictors, the bias-squared term in the MSPE can never be reduced by averaging (assuming each biased predictor has the same bias).

4.1 Best Predictor

The *best predictor* of $g(S(B))$ (with respect to squared-error loss) is

$$E[g(S(B))|Z] = \int g(t) dG_S(t|Z), \quad (4.2)$$

where

$$G_S(t|Z) = \Pr(S(B) \leq t|Z); \quad t \in \mathbb{R}.$$

Cressie (1993b) observes that if $G_S(t|Z)$ is known, then (4.2) provides a way of computing the best predictor. In Section 3.7, different methods of “estimating” $G_S(t|Z)$ (e.g., by indicator (co)kriging) were discussed, as well as the problems associated with them, such as bias in the presence of measurement error. These predictors will not be discussed further in this section.

4.2 Constrained Kriging

Following Cressie (1993b), consider predictors of $g(S(B))$ of the form $g(\mathbf{a}'Z)$. One such predictor is $g(\hat{S}_{uk}(B))$, which seems to be intuitively reasonable since $\hat{S}_{uk}(B)$ is the BLUP of $S(B)$. However, if g is nonlinear, then

$$E[g(S(B))|Z] \neq g[E(S(B)|Z)]$$

$$= g(\hat{S}_{bp}(B)).$$

Thus, even if $(S(\cdot), \epsilon(\cdot))$ were bivariate Gaussian, $g(\hat{S}_{uk}(B))$ is a biased predictor of $g(S(B))$. The reason for this bias, it turns out, is that $\hat{S}_{uk}(B)$ is *too smooth*, resulting in an often unacceptable bias (Cressie, 1993b). If one is doing many such nonlinear predictions, such as drawing a map of small areas that are above or below a prespecified threshold, this bias can result in an erroneous map.

4.2.1 Unbiasedness of $g(\mathbf{a}'\mathbf{Z})$

Assume that the coefficients \mathbf{a} satisfy the constraints,

$$(C1): E(\mathbf{a}'\mathbf{Z}) = E(S(B)),$$

$$(C2): \text{var}(\mathbf{a}'\mathbf{Z}) = \text{var}(S(B)).$$

If $(S(\cdot), \epsilon(\cdot))$ is bivariate Gaussian, and if (C1) and (C2) hold, then

$$\mathbf{a}'\mathbf{Z} \stackrel{d}{=} S(B),$$

where " $\stackrel{d}{=}$ " denotes equality in distribution. This is easy to see because Gaussian distributions are determined uniquely by their first two moments. As a consequence,

$$E[g(\mathbf{a}'\mathbf{Z})] = E[g(S(B))],$$

for any integrable function g .

If $(S(\cdot), \epsilon(\cdot))$ is not bivariate Gaussian, and if g is smooth enough to possess two derivatives, then by a suitable Taylor series approximation (δ -method), we have,

$$E[g(\mathbf{a}'\mathbf{Z})] \simeq g[E(\mathbf{a}'\mathbf{Z})] + g''[E(\mathbf{a}'\mathbf{Z})]\text{var}(\mathbf{a}'\mathbf{Z})/2,$$

and

$$E[g(S(B))] \simeq g[E(S(B))] + g''[E(S(B))]\text{var}(S(B))/2,$$

where $g''[E(\mathbf{a}'\mathbf{Z})] = \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}^2} \Big|_{\mathbf{x}=E[\mathbf{a}'\mathbf{Z}]}$, and $g''[E(S(B))] = \frac{\partial^2 g(x)}{\partial x^2} \Big|_{x=E[S(B)]}$. If, in addition, (C1) and (C2) hold, then clearly,

$$E[g(\mathbf{a}'\mathbf{Z})] \simeq E[g(S(B))],$$

for g smooth enough to possess two derivatives.

4.2.2 Optimization of $g(\mathbf{a}'\mathbf{Z})$

In light of the discussion above, we look for predictors of nonlinear functionals that are unbiased, or at least approximately so. This can be achieved for linear predictors by minimizing,

$$\text{MSPE}[g(\mathbf{a}'\mathbf{Z})],$$

with respect to \mathbf{a} , subject to constraints (C1) and (C2).

Now, if \mathbf{a} satisfies constraints (C1) and (C2), and g is smooth enough to possess two derivatives, then

$$\begin{aligned} \text{MSPE}[g(\mathbf{a}'\mathbf{Z})] &\equiv E[(g(\mathbf{a}'\mathbf{Z}) - g(S(B)))^2] \\ &\simeq \text{var}(g(\mathbf{a}'\mathbf{Z}) - g(S(B))), \\ &\simeq [g'(\mu(B))]^2 \text{MSPE}(\mathbf{a}'\mathbf{Z}), \end{aligned}$$

using the δ -method. Thus, minimizing $\text{MSPE}[g(\mathbf{a}'\mathbf{Z})]$ subject to (C1) and (C2) is approximately equivalent to minimizing $\text{MSPE}(\mathbf{a}'\mathbf{Z})$ subject to (C1) and (C2). It is this latter optimization problem that we solve, resulting in the constrained kriging predictor.

Note that the universal (or ordinary) kriging predictor is obtained by choosing \mathbf{a} to minimize $\text{MSPE}(\mathbf{a}'\mathbf{Z})$, subject only to the constraint (C1), which explains the use of the adjective “constrained” for this type of kriging (Cressie, 1993b). Also note that if g is *linear*, then

$$E[g(\mathbf{a}'\mathbf{Z})] = E[g(S(B))],$$

if and only if $E[\mathbf{a}'\mathbf{Z}] = E[S(B)]$, that is, if (C1) holds. This confirms the fact that universal (or ordinary) kriging is unbiased for the prediction of *linear* functionals of spatial processes.

4.2.3 Constrained Kriging Predictors

Assume that $\mu(B) \equiv \mathbf{x}(B)'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ and $\mathbf{x}(B)$ are defined as in (3.10) and (3.11), respectively. Then, conditions (C1) and (C2) can be written as

$$(C1): \mathbf{a}'\mathbf{X} = \mathbf{x}(B)',$$

$$(C2): \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = C(B, B),$$

since $\mathbf{a}'\mathbf{X}\boldsymbol{\beta} = E(\mathbf{a}'\mathbf{Z}) \equiv E(S(B)) = \mathbf{x}(B)'\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$, and $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \text{var}(\mathbf{a}'\mathbf{Z}) \equiv \text{var}(S(B)) = C(B, B)$.

Therefore, the objective function to minimize, with respect to \mathbf{a} , is

$$\text{MSPE}(\mathbf{a}'\mathbf{Z}) + 2\mathbf{m}_1'(\mathbf{X}'\mathbf{a} - \mathbf{x}(B)) + (m_2 - 1)(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} - C(B, B)), \quad (4.3)$$

where $2\mathbf{m}_1$ and $(m_2 - 1)$ are Lagrange multipliers (Cressie, 1993b). This yields the *universal constrained kriging predictor* of $S(B)$:

$$\hat{S}_{uc}(B) \equiv \mathbf{x}(B)'\hat{\boldsymbol{\beta}} + k_u \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (4.4)$$

where $\boldsymbol{\Sigma}$ and $\mathbf{c}(B)$ are defined as in (3.7), $\mathbf{x}(B)$ and \mathbf{X} are defined as in (3.11) and (3.12), respectively, and $\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$, defined in (3.12). Further, k_u is defined as

$$k_u \equiv \left\{ \frac{\text{var}(S(B)) - \text{var}(\mathbf{x}(B)'\hat{\boldsymbol{\beta}})}{\text{var}(\hat{S}_{uk}(B)) - \text{var}(\mathbf{x}(B)'\hat{\boldsymbol{\beta}})} \right\}^{\frac{1}{2}}, \quad (4.5)$$

where $\hat{S}_{uk}(B)$, given by (3.13), is the universal kriging predictor of $S(B)$. Consequently,

$$\text{var}(\hat{S}_{uk}(B)) = \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) + \mathbf{x}(B)'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{x}(B) - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B).$$

The MSPE of $\hat{S}_{uc}(B)$ is given by

$$\begin{aligned} \text{MSPE}(\hat{S}_{uc}(B)) &\equiv E[(\hat{S}_{uc}(B) - S(B))^2] \\ &= 2[C(B, B) - \mathbf{x}(B)'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) \\ &\quad - k_u\{\mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B)\}] \\ &= 2[C(B, B) - \mathbf{x}(B)'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) \\ &\quad - \{(C(B, B) - \mathbf{x}(B)'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{x}(B)) \times \\ &\quad (\mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B))\}^{\frac{1}{2}}]. \quad (4.6) \end{aligned}$$

If $S(\cdot)$ is first-order stationary (i.e., $\mu(s) \equiv \mu$; $s \in D$), then we call the optimal linear predictor that satisfies (C1) and (C2), the *ordinary constrained kriging predictor* of $S(B)$, written as

$$\hat{S}_{oc}(B) \equiv \hat{\mu} + k_o \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1}), \quad (4.7)$$

where $\hat{\mu} \equiv \mathbf{1}'\Sigma^{-1}\mathbf{Z}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$ is the BLUE of μ , k_o is defined as

$$k_o \equiv \left\{ \frac{\text{var}(S(B)) - \text{var}(\hat{\mu})}{\text{var}(\hat{S}_{ok}(B)) - \text{var}(\hat{\mu})} \right\}^{\frac{1}{2}}, \quad (4.8)$$

and $\hat{S}_{ok}(B)$ is the ordinary kriging predictor of $S(B)$ given by (3.15).

Just as for universal constrained kriging, the MSPE of $\hat{S}_{oc}(B)$ is given by

$$\begin{aligned} \text{MSPE}(\hat{S}_{oc}(B)) &\equiv E[(\hat{S}_{oc}(B) - S(B))^2] \\ &= 2[C(B, B) - (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}'\Sigma^{-1}\mathbf{c}(B) \\ &\quad - k_o\{\mathbf{c}(B)'\Sigma^{-1}\mathbf{c}(B) - (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}(\mathbf{1}'\Sigma^{-1}\mathbf{c}(B))^2\}] \\ &= 2[C(B, B) - (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}'\Sigma^{-1}\mathbf{c}(B) \\ &\quad - \{(C(B, B) - (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}) \times \\ &\quad (\mathbf{c}(B)'\Sigma^{-1}\mathbf{c}(B) - (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}(\mathbf{1}'\Sigma^{-1}\mathbf{c}(B))^2)\}^{\frac{1}{2}}]. \end{aligned} \quad (4.9)$$

4.2.4 Existence of the Constrained Kriging Predictor

Examination of equation (4.5) indicates that $\hat{S}_{uc}(B)$ exists only if

$$(E1) : \text{var}(\hat{S}_{uk}(B)) > \text{var}(\mathbf{x}(B)'\hat{\beta}),$$

$$(E2) : \text{var}(S(B)) \geq \text{var}(\mathbf{x}(B)'\hat{\beta}).$$

If $S(\cdot)$ is first-order stationary (i.e., $\mu(s) \equiv \mu$; $s \in D$), then examination of equation (4.8) indicates that $\hat{S}_{oc}(B)$ exists only if

$$(E1) : \text{var}(\hat{S}_{ok}(B)) > \text{var}(\hat{\mu}),$$

$$(E2) : \text{var}(S(B)) \geq \text{var}(\hat{\mu}).$$

Proposition 4.1 *For the variances, $\text{var}(\hat{S}_{uk}(B))$ and $\text{var}(\mathbf{x}(B)'\hat{\beta})$, in (E1), the following inequality holds,*

$$\text{var}(\hat{S}_{uk}(B)) \geq \text{var}(\mathbf{x}(B)'\hat{\beta}).$$

Proof: Since $\text{cov}(\hat{\beta}, (\mathbf{Z} - \mathbf{X}\hat{\beta})) = 0$, $\text{var}(\hat{S}_{uk}(B)) = \text{var}(\mathbf{x}(B)'\hat{\beta} + \mathbf{c}(B)'\Sigma^{-1}(\mathbf{Z} - \mathbf{X}\hat{\beta})) = \text{var}(\mathbf{x}(B)'\hat{\beta}) + \text{var}(\mathbf{c}(B)'\Sigma^{-1}(\mathbf{Z} - \mathbf{X}\hat{\beta})) \geq \text{var}(\mathbf{x}(B)'\hat{\beta})$. \square

Thus, existence condition (E1) can only be violated if

$$\text{var}(\hat{S}_{uk}(B)) = \text{var}(\mathbf{x}(B)' \hat{\beta}). \quad (4.10)$$

Proposition 4.2 Let $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))'$ be a data vector. Suppose that $\text{var}(\mathbf{Z}) = \Sigma$ is positive definite (p.d.). Then,

$$\text{var}(\hat{\mu}) = (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \leq \min_{i=1, \dots, n} \{\text{var}(Z(s_i))\},$$

where $\hat{\mu} = (\mathbf{1}' \Sigma^{-1} \mathbf{Z}) / (\mathbf{1}' \Sigma^{-1} \mathbf{1})$.

Proof: Rearrange the spatial labels $\{s_1, \dots, s_n\}$ of \mathbf{Z} so that $\text{var}(Z(s_1)) = \min_{i=1, \dots, n} \{\text{var}(Z(s_i))\}$.

Then we can write

$$\Sigma \equiv \begin{bmatrix} \sigma_1^2 & \mathbf{c}_1' \\ \mathbf{c}_1 & \Sigma_1 \end{bmatrix},$$

where $\sigma_1^2 = \text{var}(Z(s_1))$, $\mathbf{c}_1 = (C(s_1, s_2), \dots, C(s_1, s_n))'$, and $\Sigma_1 = \text{var}(Z(s_2), \dots, Z(s_n))$. Define

$$\mathbf{D} \equiv \Sigma_1 - \sigma_1^{-2} \mathbf{c}_1 \mathbf{c}_1';$$

then it is elementary to show that \mathbf{D} is p.d. Notice that (e.g., Mardia et al. 1979, App. A.2.4),

$$\Sigma^{-1} = \begin{bmatrix} \sigma_1^{-2} + \sigma_1^{-4} \mathbf{c}_1' \mathbf{D}^{-1} \mathbf{c}_1 & -\sigma_1^{-2} \mathbf{c}_1' \mathbf{D}^{-1} \\ -\sigma_1^{-2} \mathbf{D}^{-1} \mathbf{c}_1 & \mathbf{D}^{-1} \end{bmatrix}.$$

Thus,

$$\begin{aligned} (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} &= (\sigma_1^{-2} + \sigma_1^{-4} \mathbf{c}_1' \mathbf{D}^{-1} \mathbf{c}_1 - 2\sigma_1^{-2} \mathbf{c}_1' \mathbf{D}^{-1} \mathbf{1} + \mathbf{1}' \mathbf{D}^{-1} \mathbf{1})^{-1} \\ &= (\sigma_1^{-2} + (\sigma_1^{-2} \mathbf{c}_1 - \mathbf{1})' \mathbf{D}^{-1} (\sigma_1^{-2} \mathbf{c}_1 - \mathbf{1}))^{-1} \\ &\leq \sigma_1^2, \end{aligned}$$

since \mathbf{D}^{-1} is p.d. If $\sigma_1^{-2} \mathbf{c}_1 \neq \mathbf{1}$, then $(\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} < \sigma_1^2 = \min_{i=1, \dots, n} \{\text{var}(Z(s_i))\}$. \square

Thus, if $S(\cdot)$ is first-order stationary, then existence condition (E2) cannot be violated if

$$\text{var}(S(B)) \geq \min_{i=1, \dots, n} \{\text{var}(Z(s_i))\}. \quad (4.11)$$

For example, for $B = \{s_0\}$, $\tau^2 = 0$, and $\text{var}(S(s)) \equiv \sigma^2$; $s \in D$, the existence condition (E2) is never violated.

4.2.5 Geometric Interpretation of Constrained Kriging

It is sometimes useful to explore geometric properties of mathematical/statistical quantities that are based on vector spaces. This section explores the geometric properties of constrained kriging, and for ease of presentation, we shall assume throughout this section that $S(\cdot)$ is second-order stationary. That is, assume that

$$\begin{aligned} E(S(s)) &= \mu; \quad s \in D, \\ C(s, u) &= C^o(s - u); \quad s, u \in D, \end{aligned}$$

and denote $\text{var}(S(s)) = C^o(0) \equiv \sigma^2$; $s \in D$.

Consider the ordinary constrained kriging predictor $\hat{S}_{oc}(B)$ of $S(B)$; $B \in D$. Recall that

$$\hat{S}_{oc}(B) = \mathbf{a}'_0 \mathbf{Z},$$

where \mathbf{a}_0 minimizes, with respect to \mathbf{a} ,

$$\begin{aligned} \text{MSPE}(\mathbf{a}'\mathbf{Z}) &\equiv E[(\mathbf{a}'\mathbf{Z} - S(B))^2] \\ &= C(B, B) + \mathbf{a}'\Sigma\mathbf{a} - 2\mathbf{a}'\mathbf{c}(B), \end{aligned}$$

subject to

$$(C1): \quad \mathbf{a}'\mathbf{1} = 1,$$

$$(C2): \quad \mathbf{a}'\Sigma\mathbf{a} = C(B, B).$$

These are special cases of (C1) and (C2) given in Section 4.2.3, since $\mathbf{a}'\mathbf{1}\mu = E(\mathbf{a}'\mathbf{Z}) \equiv E(S(B)) = \mu$, for all μ , and $\mathbf{a}'\Sigma\mathbf{a} = \text{var}(\mathbf{a}'\mathbf{Z}) \equiv \text{var}(S(B)) = C(B, B)$.

Geometric Interpretation of the Constraints (C1) and (C2)

Define

$$\begin{aligned} \mathcal{A} &\equiv \{\mathbf{a} : (C1) \text{ and } (C2) \text{ hold}\} \\ &= \{\mathbf{a} : \mathbf{a}'\mathbf{1} = 1, \mathbf{a}'\Sigma\mathbf{a} = C(B, B)\}. \end{aligned} \tag{4.12}$$

Observe that if \mathcal{A} is not ϕ , the empty set, then $\mathbf{a}_0 \in \mathcal{A}$.

Proposition 4.3 *Constraint (C1) describes an $(n-1)$ -dimensional hyperplane in \mathbb{R}^n , which contains the orthonormal basis vectors, $\mathbf{e}_1 = (1, 0, \dots, 0)'$, \dots , $\mathbf{e}_n = (0, \dots, 0, 1)'$.*

Proof: Clearly, the equation $\mathbf{a}'\mathbf{1} = 1$; $\mathbf{a} \in \mathbb{R}^n$, represents an $(n-1)$ -dimensional hyperplane in \mathbb{R}^n . Further, $\mathbf{e}_i'\mathbf{1} = 1$; $i = 1, \dots, n$, demonstrates that the hyperplane contains $\mathbf{e}_1, \dots, \mathbf{e}_n$. \square

Proposition 4.4 *Assume that Σ is p.d. Then constraint (C2) describes an $(n-1)$ -dimensional ellipsoid in \mathbb{R}^n , centered on the origin and containing the vectors $q\mathbf{e}_1, \dots, q\mathbf{e}_n$, where*

$$q = \{\text{var}(S(B))/\text{var}(Z(s))\}^{\frac{1}{2}} = \{C(B, B)/(\sigma^2 + \tau^2)\}^{\frac{1}{2}}. \quad (4.13)$$

If $C^o(\mathbf{h}) = 0$; $\mathbf{h} \neq 0$, then (C2) describes an $(n-1)$ -dimensional sphere in \mathbb{R}^n .

Proof: If Σ is p.d., then the equation $\mathbf{a}'\Sigma\mathbf{a} = C(B, B)$ can be written as

$$\sum_{i=1}^n a_i^2 + \frac{2\sigma^2}{\sigma^2 + \tau^2} \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \rho^o(\mathbf{s}_i - \mathbf{s}_j) = C(B, B)/(\sigma^2 + \tau^2), \quad (4.14)$$

where $\rho^o(\mathbf{s}_i - \mathbf{s}_j) = C^o(\mathbf{s}_i - \mathbf{s}_j)/\sigma^2$. Clearly, (4.14) describes an $(n-1)$ -dimensional ellipsoid in \mathbb{R}^n , centered on the origin, with the axes of symmetry not necessarily equal to the basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$.

Now, if we put $\mathbf{a} = q\mathbf{e}_i$, we obtain

$$q^2 \mathbf{e}_i' \Sigma \mathbf{e}_i = q^2(\sigma^2 + \tau^2) = C(B, B); \quad i = 1, \dots, n,$$

where $\mathbf{e}_1, \dots, \mathbf{e}_n$ are the orthonormal basis vectors, and q is given by (4.13). Therefore, the ellipsoid contains the vectors $q\mathbf{e}_1, \dots, q\mathbf{e}_n$.

If $C^o(\mathbf{h}) = 0$; $\mathbf{h} \neq 0$, then $\mathbf{a}'\Sigma\mathbf{a} = (\sigma^2 + \tau^2)\mathbf{a}'\mathbf{a}$. Hence, (4.14) reduces to

$$\mathbf{a}'\mathbf{a} = C(B, B)/(\sigma^2 + \tau^2),$$

which describes an $(n-1)$ -dimensional sphere in \mathbb{R}^n . \square

Recall that here, $S(\cdot)$ is assumed to be second-order stationary. Then, by Cartier's formula (Section 3),

$$C(B, B) \equiv \text{var}(S(B)) \leq \text{var}(S(s_0)) \equiv \sigma^2.$$

Thus, if $\tau^2 = 0$ and $B = \{s_0\}$, then $\text{var}(S(B)) = \text{var}(Z(s))$, and $q = 1$. If $\tau^2 > 0$, then $\text{var}(S(B)) \leq \text{var}(S(s)) < \text{var}(Z(s))$, and $q < 1$. If B contains more than one element and $\text{var}(S(B)) < \text{var}(S(s_0))$, then $\text{var}(S(B)) < \text{var}(Z(s))$, whatever τ^2 is, and hence $q < 1$.

Proposition 4.5 *Recall the definition of \mathcal{A} from (4.12). If \mathcal{A} has more than one element, then \mathcal{A} describes an $(n-2)$ -dimensional ellipsoid in \mathbb{R}^n ; $n \geq 3$. This $(n-2)$ -dimensional ellipsoid is the intersection set between the $(n-1)$ -dimensional hyperplane described by (C1) and the $(n-1)$ -dimensional ellipsoid described by (C2). For $n = 2$, if $\mathcal{A} \neq \emptyset$, then \mathcal{A} contains no more than two elements.*

Proof: The set $\mathcal{A} = \{a : (C1) \text{ and } (C2) \text{ hold}\}$ is clearly the intersection set between the $(n-1)$ -dimensional hyperplane described by (C1) and the $(n-1)$ -dimensional ellipsoid described by (C2). If $n \geq 3$, then clearly, \mathcal{A} may be empty, it may contain one element, or it may contain the elements of a set describing an ellipsoid with one dimension less than that described by (C2). If $n = 2$, then (C1) describes a line and (C2) describes an ellipse in \mathbb{R}^2 . Clearly, \mathcal{A} may be empty, it may contain one element, or it may contain two elements. \square

Note that if $q = 1$, then the $(n-2)$ -dimensional ellipsoid described by \mathcal{A} contains the vectors e_1, \dots, e_n , since e_1, \dots, e_n belong to the sets described by both (C1) and (C2). For example, when $n = 3$, there is an ellipsoid that passes through the value of unity on each of the three orthogonal axes.

Proposition 4.6 *Assume that $\mathcal{A} \neq \emptyset$. Then, if $a \in \mathcal{A}$,*

$$\text{MSPE}(a'Z) = 2(C(B, B) - a'c(B)); \quad (4.15)$$

that is, $\text{MSPE}(a'Z)$ is linear in a , for all $a \in \mathcal{A}$.

Proof: If $\mathbf{a} \in \mathcal{A}$, then \mathbf{a} satisfies (C1) and (C2). Thus,

$$\begin{aligned}\text{MSPE}(\mathbf{a}'\mathbf{Z}) &= \mathbf{a}'\Sigma\mathbf{a} + C(B, B) - 2\mathbf{a}'\mathbf{c}(B) \\ &= 2(C(B, B) - \mathbf{a}'\mathbf{c}(B)),\end{aligned}$$

since $\mathbf{a}'\Sigma\mathbf{a} = C(B, B)$. □

To illustrate these geometric properties, consider a simple example where $n = 2$, $B = \{\mathbf{s}_0\}$, and $\tau^2 = 0$, in which case $q = 1$; see Figure 4.1. Notice the following features:

1. Constraint (C1) describes a 1-dimensional plane (line) in \mathbb{R}^2 , containing the orthonormal basis vectors, $\mathbf{e}_1 = (1, 0)'$ and $\mathbf{e}_2 = (0, 1)'$.

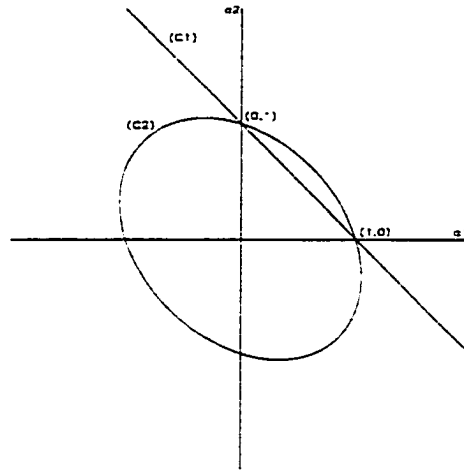


Figure 4.1 Geometric interpretation of constraints (C1) and (C2), when $n = 2$, $B = \{\mathbf{s}_0\}$, and $\tau^2 = 0$.

2. Constraint (C2) describes a 1-dimensional ellipsoid (ellipse) in \mathbb{R}^2 , centered on the origin and containing the vectors \mathbf{e}_1 and \mathbf{e}_2 (since $B = \{\mathbf{s}_0\}$ and $\tau^2 = 0$). If $C^o(\mathbf{s}_i - \mathbf{s}_j) = 0$; $i \neq j$, then (C2) describes a 1-dimensional sphere (circle) in \mathbb{R}^2 .
3. The set $\mathcal{A} = \{\mathbf{e}_1, \mathbf{e}_2\} \neq \emptyset$. If, in this example of $B = \{\mathbf{s}_0\}$ and $\tau^2 = 0$, we were to increase the number of data to $n = 3$, then \mathcal{A} would describe a 1-dimensional ellipsoid

(ellipse) in \mathbb{R}^3 , containing the orthonormal basis vectors, $\mathbf{e}_1 = (1, 0, 0)'$, $\mathbf{e}_2 = (0, 1, 0)'$, and $\mathbf{e}_3 = (0, 0, 1)'$.

4. Recall from Proposition 4.6, that the function $\text{MSPE}(\mathbf{a}'\mathbf{Z})$ is *linear* in \mathbf{a} , for all $\mathbf{a} \in \mathcal{A}$; that is, in this example,

$$\text{MSPE}(\mathbf{a}'\mathbf{Z}) = 2(\sigma^2 - \mathbf{a}'\mathbf{c}(\mathbf{s}_0)),$$

where $\sigma^2 = \text{var}(S(\mathbf{s}_0))$ and $\mathbf{c}(\mathbf{s}_0) = (C^o(\mathbf{s}_1 - \mathbf{s}_0), C^o(\mathbf{s}_2 - \mathbf{s}_0))'$. Then $C^o(\mathbf{s}_1 - \mathbf{s}_0) > C^o(\mathbf{s}_2 - \mathbf{s}_0)$ implies that

$$\begin{aligned} \text{MSPE}(\mathbf{e}_1'\mathbf{Z}) &= 2(\sigma^2 - C^o(\mathbf{s}_1 - \mathbf{s}_0)) \\ &< 2(\sigma^2 - C^o(\mathbf{s}_2 - \mathbf{s}_0)) \\ &= \text{MSPE}(\mathbf{e}_2'\mathbf{Z}). \end{aligned}$$

Therefore, $C^o(\mathbf{s}_1 - \mathbf{s}_0) > C^o(\mathbf{s}_2 - \mathbf{s}_0)$ implies that $\mathbf{a}_0 = \mathbf{e}_1$, and

$$\hat{S}_{oc}(\mathbf{s}_0) = Z(\mathbf{s}_1).$$

Geometric Interpretation of the Existence Conditions (E1) and (E2)

Recall that in this section $S(\cdot)$ is assumed to be second-order stationary. For the remainder of this section, also assume that Σ is p.d.

Existence condition (E1): Assume $\mathcal{A} \neq \emptyset$. From (4.10), the condition (E1) is violated if and only if

$$\begin{aligned} \text{var}(\hat{S}_{ok}(B)) &= (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1} + \mathbf{c}(B)'\Sigma^{-1}\mathbf{c}(B) - (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}(\mathbf{c}(B)'\Sigma^{-1}\mathbf{1})^2 \\ &\equiv (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1} = \text{var}(\hat{\mu}), \end{aligned}$$

that is, if and only if

$$\mathbf{c}(B)'\Sigma^{-1}\mathbf{c}(B) = (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}(\mathbf{c}(B)'\Sigma^{-1}\mathbf{1})^2. \quad (4.16)$$

Since Σ is p.d., there exists a nonsingular square matrix R , such that $\Sigma = RR'$. Then, by the Cauchy-Schwarz inequality, (4.16) occurs if and only if $\mathbf{c}(B)'R = 0$, $\mathbf{1}'R = 0$, or

$\mathbf{c}(B)'R = \alpha \mathbf{1}'R$, $\alpha \in \mathbb{R}$. Clearly, Σ p.d. implies that $\mathbf{1}'R \neq 0$; therefore (4.16) holds if and only if

$$\mathbf{c}(B) = \alpha \mathbf{1}; \quad \alpha \in \mathbb{R},$$

since R is nonsingular. Note that this includes the case $\mathbf{c}(B) = \mathbf{0}$, since we may put $\alpha = 0$.

As a consequence,

$$\text{MSPE}(\mathbf{a}'\mathbf{Z}) = 2(C(B, B) - \alpha \mathbf{a}'\mathbf{1}) = 2(C(B, B) - \alpha); \quad \alpha \in \mathbb{R}.$$

Thus, the violation of condition (E1) implies that $\text{MSPE}(\mathbf{a}'\mathbf{Z})$ is constant in \mathbf{a} , for all $\mathbf{a} \in \mathcal{A}$. This means that $\text{MSPE}(\mathbf{a}'\mathbf{Z})$ is minimized for any $\mathbf{a} \in \mathcal{A}$, and no *unique* solution exists.

Now, $\mathbf{c}(B) = \alpha \mathbf{1}$; $\alpha \in \mathbb{R}$, occurs when $C(\mathbf{s}_1, B) = \dots = C(\mathbf{s}_n, B) = \alpha \in \mathbb{R}$. If $\alpha \neq 0$, this occurs when the “distance” (as defined by the semivariogram “metric” $\gamma(\cdot, \cdot)$) between the location of the predictand and the location of each datum is equal. A simple practical solution to this problem may be to shift the prediction location slightly closer (according to the metric $\gamma(\cdot, \cdot)$) to one of the data, if possible.

Existence condition (E2): This condition is violated if and only if

$$\text{var}(S(B)) = C(B, B) \equiv \mathbf{a}'\Sigma\mathbf{a} < (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1} = \text{var}(\hat{\mu}),$$

that is, if and only if

$$(\mathbf{a}'\Sigma\mathbf{a})(\mathbf{1}'\Sigma^{-1}\mathbf{1}) < 1, \quad \text{for all } \mathbf{a} \in \mathcal{A}.$$

Since Σ is p.d., there exists a nonsingular matrix \mathbf{R} , such that $\Sigma = \mathbf{R}\mathbf{R}'$. Then, by the Cauchy-Schwarz inequality,

$$\begin{aligned} (\mathbf{a}'\mathbf{1})^2 &= (\mathbf{a}'\mathbf{R}\mathbf{R}^{-1}\mathbf{1})^2 \\ &\leq (\mathbf{a}'\mathbf{R}\mathbf{R}'\mathbf{a})(\mathbf{1}'(\mathbf{R}^{-1})'\mathbf{R}^{-1}\mathbf{1}) \\ &= (\mathbf{a}'\Sigma\mathbf{a})(\mathbf{1}'\Sigma^{-1}\mathbf{1}). \end{aligned}$$

Combining these two consequences, we see that

$$\text{var}(S(B)) < \text{var}(\hat{\mu})$$

implies that

$$a'1 < 1,$$

which contradicts (C1). Hence

$$\mathcal{A} = \phi,$$

and no solution exists. This is in contrast to the violation of (E1), where one has a solution but it is not unique. See Figure 4.2 for an illustration of violation of (E2) in the case $n = 2$.

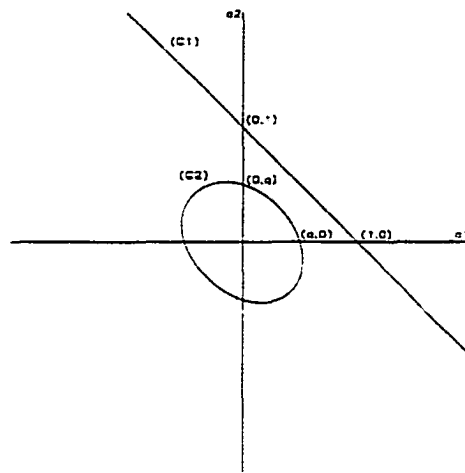


Figure 4.2 Geometric interpretation of the violation of existence condition (E2), when $n = 2$.

If (E2) is violated, there does not appear to be an obvious way around the problem, except to increase the sample size n sufficiently to decrease $\text{var}(\hat{\mu})$, or to decrease the aggregation size of B sufficiently to increase $\text{var}(S(B))$, enough for (E2) to hold.

4.2.6 Affine Correction

It has been pointed out that universal (or ordinary) kriging is too smooth to be useful as a predictor of nonlinear functionals of spatial processes (e.g., Journel and Huijbregts, 1978, Ch. VI, and Cressie, 1993b). There exists a simple “affine-correction” procedure that can

“roughen” (increase the variability) of predictors that are too smooth. The universal (or ordinary) constrained kriging predictor is an affine-corrected universal (or ordinary) kriging predictor, but we show that the way the affine-correction procedure is applied in this case, differs from the way the procedure has usually been applied. Therefore, past conclusions made about traditionally affinely corrected predictors are strictly speaking not applicable to constrained kriging predictors.

The affine-correction procedure may be described as follows: Suppose that we have two random variables, X and Y , where

$$E(X) = \mu_x \quad \text{and} \quad \text{var}(X) = \sigma_x^2,$$

$$E(Y) = \mu_y \quad \text{and} \quad \text{var}(Y) = \sigma_y^2,$$

and we wish to “correct” X so that its mean and variance match those of Y . Assume that μ_x, μ_y, σ_x^2 , and σ_y^2 are known, and define the random variable X' as

$$X' = \mu_y + \frac{\sigma_y}{\sigma_x}(X - \mu_x). \quad (4.17)$$

Then

$$E(X') = \mu_y \quad \text{and} \quad \text{var}(X') = \sigma_y^2,$$

so that X' has the desired distributional properties.

If μ_x and μ_y are unknown, then X' is not a statistic. Now suppose that unbiased estimators $\hat{\mu}_x$ and $\hat{\mu}_y$ of μ_x and μ_y , respectively, are available, where $\text{var}(\hat{\mu}_x) = \tau_x^2$ and $\text{var}(\hat{\mu}_y) = \tau_y^2$.

Proposition 4.7 *Define*

$$X'' = \hat{\mu}_y + \left(\frac{\sigma_y^2 - \tau_y^2}{\text{var}(X - \hat{\mu}_x)} \right)^{\frac{1}{2}} (X - \hat{\mu}_x), \quad (4.18)$$

where $\hat{\mu}_x$ and $\hat{\mu}_y$ are unbiased estimators of μ_x and μ_y , respectively, and $\text{var}(\hat{\mu}_x) = \tau_x^2$ and $\text{var}(\hat{\mu}_y) = \tau_y^2$. If $\text{cov}((X - \hat{\mu}_x), \hat{\mu}_y) = 0$, then

$$E(X'') = \mu_y \quad \text{and} \quad \text{var}(X'') = \sigma_y^2.$$

Proof: The estimator $\hat{\mu}_x$ of μ_x is unbiased; therefore, $E(\hat{\mu}_x) = \mu_x = E(X)$. Thus,

$$E(X'') = E(\hat{\mu}_y) = \mu_y,$$

since $\hat{\mu}_y$ is an unbiased estimator of μ_y .

In addition, $\text{cov}((X - \hat{\mu}_x), \hat{\mu}_y) = 0$ implies that

$$\begin{aligned} \text{var}(X'') &= \text{var}(\hat{\mu}_y) + \left(\frac{\sigma_y^2 - \tau_y^2}{\text{var}(X - \hat{\mu}_x)} \right) \text{var}(X - \hat{\mu}_x) \\ &= \text{var}(\hat{\mu}_y) + \sigma_y^2 - \tau_y^2 = \sigma_y^2, \end{aligned}$$

since $\text{var}(\hat{\mu}_y) = \tau_y^2$. □

In geostatistical applications, the affine-correction given by equation (4.17) has been used (e.g., Journel and Huijbregts, 1978, Ch. VI, and Lantuejoul, 1988), even if the means are unknown; estimators are substituted for the unknown values but no compensation of the sort given by (4.18) is made.

Suppose we wish to predict $S(B)$, where

$$E(S(B)) = \mathbf{x}(B)' \boldsymbol{\beta} \quad \text{and} \quad \text{var}(S(B)) = C(B, B),$$

where $\mathbf{x}(B)$ and $\boldsymbol{\beta}$ are defined as in (3.10) and (3.11), and $C(B, B) \equiv \frac{1}{|B|^2} \int_B \int_B C(\mathbf{s}, \mathbf{u}) \, d\mathbf{s} \, d\mathbf{u}$. Consider the simple kriging predictor of $S(B)$,

$$\hat{S}_{sk}(B) \equiv \mathbf{x}(B)' \boldsymbol{\beta} + \mathbf{c}(B)' \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X} \boldsymbol{\beta}),$$

where $\mathbf{c}(B)$, $\boldsymbol{\Sigma}$, and \mathbf{X} are defined as in (3.7). Then

$$E(\hat{S}_{sk}(B)) = \mathbf{x}(B)' \boldsymbol{\beta},$$

but

$$\text{var}(\hat{S}_{sk}(B)) = \mathbf{c}(B)' \boldsymbol{\Sigma} \mathbf{c}(B),$$

which, in general, is not equal to $C(B, B) = \text{var}(S(B))$. Therefore, an affine-corrected simple kriging predictor is defined as

$$\hat{S}_{as}(B) \equiv \mathbf{x}(B)' \boldsymbol{\beta} + \left(\frac{C(B, B)}{\mathbf{c}(B)' \boldsymbol{\Sigma} \mathbf{c}(B)} \right)^{\frac{1}{2}} (\hat{S}_{sk}(B) - \mathbf{x}(B)' \boldsymbol{\beta}).$$

Observe that the application of the procedure illustrated above depends on the assumption that β and the parameters that govern $\{\text{var}(S(B)), \text{var}(\hat{S}_{sk}(B))\}$ are all known. Assume now that β is unknown, in precisely the same way as one does for universal kriging. In this case, $\hat{S}_{as}(B)$ is no longer a statistic. The parameter vector β may then be estimated by the generalized-least-squares estimator,

$$\hat{\beta} \equiv (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Z,$$

which is the BLUE of β . Then, $S(B)$ may be predicted by the BLUP, which is the universal kriging predictor

$$\hat{S}_{uk}(B) \equiv x(B)' \hat{\beta} + c(B)' \Sigma^{-1} (Z - X \hat{\beta}),$$

where

$$E(\hat{S}_{uk}(B)) = x(B)' \beta.$$

Although $\hat{S}_{uk}(B)$ is unbiased, it is usually much smoother than $S(B)$. It is easy to see that,

$$\text{var}(\hat{S}_{uk}(B)) = x(B)' (X' \Sigma^{-1} X)^{-1} x(B) + c(B)' \Sigma c(B) - c(B)' \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} c(B),$$

which, in general, is not equal to $C(B, B) = \text{var}(S(B))$. Now, $\text{cov}((Z - X \hat{\beta}), \hat{\beta}) = 0$, and hence an affine-corrected universal kriging predictor is defined as

$$\begin{aligned} \hat{S}_{au}(B) &\equiv x(B)' \beta + \left(\frac{C(B, B) - \text{var}(x(B)' \hat{\beta})}{\text{var}(\hat{S}_{uk}(B) - x(B)' \hat{\beta})} \right)^{\frac{1}{2}} (\hat{S}_{uk}(B) - x(B)' \hat{\beta}) \\ &= \hat{S}_{uc}(B), \end{aligned}$$

where $\hat{S}_{uc}(B)$ is the universal constrained kriging predictor given by (4.4). This is because $\text{cov}((Z - X \hat{\beta}), \hat{\beta}) = 0$ implies that

$$\text{var}(\hat{S}_{uk}(B) - x(B)' \hat{\beta}) = \text{var}(\hat{S}_{uk}(B)) - \text{var}(x(B)' \hat{\beta}).$$

Therefore, in conclusion, the constrained kriging predictor can also be interpreted as an affinely corrected kriging predictor, where the variability in estimating mean parameters is directly, and correctly, accounted for.

4.3 Constrained Kriging and Uncorrelated Spatial Processes

Suppose that the signal process $S(\cdot)$ is spatially uncorrelated. Then we saw in Section 3.5 that the kriging predictor is equal to the estimated mean (i.e., the BLUE of $\mu(\cdot)$), if the prediction location is different from the sampling location. However, not surprisingly, we see that constrained kriging behaves differently than kriging for spatially uncorrelated processes. For example, the ordinary constrained kriging predictor is “pulled” towards the datum nearest the prediction location, and the strength of this “pull” depends upon the measurement error. If no measurement error exists, then the ordinary constrained kriging predictor is equal to the nearest datum.

Define $A^* = B \cap A$, where $B = \{\mathbf{u}_1, \dots, \mathbf{u}_N\} \subset D$ is the set of prediction locations, and $A = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ is the sample.

Proposition 4.8 *Suppose that $S(\cdot)$ is a spatially uncorrelated process. If $A^* \neq \emptyset$, then the universal constrained kriging predictor of $S(B)$ is*

$$\hat{S}_{uc}(B) = \mathbf{x}(B)' \hat{\beta} + k_u \frac{1}{N} \sum_{i=1}^n C(\mathbf{s}_i, \mathbf{s}_i) I(\mathbf{s}_i \in A^*) \mathbf{e}_i' \Sigma^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}), \quad (4.19)$$

where

$$\begin{aligned} k_u &= \left\{ \frac{\text{var}(S(B)) - \text{var}(\mathbf{x}(B)' \hat{\beta})}{\text{var}(\hat{S}_{uc}(B)) - \text{var}(\mathbf{x}(B)' \hat{\beta})} \right\}^{\frac{1}{2}} \\ &= \left\{ \frac{C(B, B) - \mathbf{x}(B)' (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{x}(B)}{\mathbf{c}(B)' \Sigma \mathbf{c}(B) - \mathbf{c}(B)' \Sigma^{-1} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{c}(B)} \right\}^{\frac{1}{2}}, \end{aligned}$$

and the ordinary constrained kriging predictor of $S(B)$ is

$$\hat{S}_{oc}(B) = \hat{\mu} + k_o \frac{1}{N} \sum_{i=1}^n C(\mathbf{s}_i, \mathbf{s}_i) I(\mathbf{s}_i \in A^*) \mathbf{e}_i' \Sigma^{-1} (\mathbf{Z} - \hat{\mu} \mathbf{1}), \quad (4.20)$$

where

$$\begin{aligned} k_o &= \left\{ \frac{\text{var}(S(B)) - \text{var}(\hat{\mu})}{\text{var}(\hat{S}_{oc}(B)) - \text{var}(\hat{\mu})} \right\}^{\frac{1}{2}} \\ &= \left\{ \frac{C(B, B) - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1}}{\mathbf{c}(B)' \Sigma \mathbf{c}(B) - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} (\mathbf{c}(B)' \Sigma^{-1} \mathbf{1})^2} \right\}^{\frac{1}{2}}, \end{aligned}$$

with $\mathbf{c}(B)$ and Σ defined as in (3.23) and (3.24), respectively.

Proof: If $A^* \neq \phi$, then

$$c(B) = \frac{1}{N} \sum_{i=1}^n C(s_i, s_i) I(s_i \in A^*) \mathbf{e}_i \neq \mathbf{0},$$

and the result follows. \square

If $A^* = B \cap A = \phi$, then the denominator of k_u or k_o is zero, and so a solution to the constrained kriging equations cannot be found directly. However, a solution can be found indirectly by examining the limiting results of the predictors as the strength of spatial correlation tends to zero (see Aldworth and Cressie, 1998; App. 1, and the proposition below).

Define $A^\# = \{\arg \min_{s \in A} \{\inf_{u \in B} \|s - u\|\}\}$, and denote $A^\# \equiv \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, where $m = |A^*| \leq n$. Note that if $A^* \neq \phi$, then $A^* = A^\#$.

Proposition 4.9 *Suppose that $S(\cdot)$ is a spatially uncorrelated process, and suppose that $A^* = \phi$. Then: (a) If $\mu(B) \equiv \mathbf{x}(B)' \boldsymbol{\beta}$ and $\text{var}(S(\mathbf{s})) \equiv \sigma^2$, then the universal constrained kriging predictor of $S(B)$ is*

$$\begin{aligned} \hat{S}_{uc}(B) = & \mathbf{x}(B)' \hat{\boldsymbol{\beta}} + \left\{ \frac{\sigma^2 - N(\sigma^2 + \tau^2) \mathbf{x}(B)' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}(B)}{N(\sigma^2 + \tau^2) \left\{ \sum_{i=1}^m N_i^2 - \sum_{i=1}^m \sum_{j=1}^m N_i N_j \mathbf{x}(\mathbf{v}_i)' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}(\mathbf{v}_j) \right\}} \right\}^{\frac{1}{2}} \\ & \times \left(\sum_{i=1}^m N_i (Z(\mathbf{v}_i) - \mathbf{x}(\mathbf{v}_i)' \hat{\boldsymbol{\beta}}) \right); \end{aligned} \quad (4.21)$$

(b) If $S(\cdot)$ is second-order stationary, then the ordinary constrained kriging predictor of $S(B)$ is

$$\hat{S}_{oc}(B) = \bar{Z} + \left\{ \frac{n\sigma^2 - N(\sigma^2 + \tau^2)}{N(\sigma^2 + \tau^2) \{n \sum_{i=1}^m N_i^2 - N_m^2\}} \right\}^{\frac{1}{2}} \left(\sum_{i=1}^m N_i Z(\mathbf{v}_i) - N_m \bar{Z} \right), \quad (4.22)$$

where $N = |B|$, $N_i \equiv |\{u \in B : \|u - \mathbf{v}_i\| = \min_{j=1, \dots, |B|} \|u_j - \mathbf{v}_i\|\}|$, $i = 1, \dots, m$, and $N_m = \sum_{i=1}^m N_i$.

Proof: Part (a) is easy to prove, following the method of Aldworth and Cressie (1998; App. 1). Part (b) is specifically proved in Aldworth and Cressie (1998), although it is easily seen to follow as a special case of part (a) where $\mathbf{x}(\mathbf{s})$ is a single element equal to 1. \square

Several consequences of Proposition 4.8 and Proposition 4.9 follow:

1. Suppose $B = \{s_0\}$, where $s_0 \notin A$. Then it follows that (4.21) in Proposition 4.9 reduces to

$$\begin{aligned} \hat{S}_{uc}(s_0) = & \mathbf{x}(s_0)' \hat{\beta} + \left\{ \frac{\sigma^2 - (\sigma^2 + \tau^2) \mathbf{x}(s_0)' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}(s_0)}{(\sigma^2 + \tau^2) \{m - \sum_{i=1}^m \sum_{j=1}^m \mathbf{x}(\mathbf{v}_i)' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}(\mathbf{v}_j)\}} \right\}^{\frac{1}{2}} \\ & \times \left(\sum_{i=1}^m (Z(\mathbf{v}_i) - \mathbf{x}(\mathbf{v}_i)' \hat{\beta}) \right), \end{aligned} \quad (4.23)$$

and (4.22) in Proposition 4.9 reduces to

$$\hat{S}_{oc}(s_0) = \bar{Z} + \left\{ \frac{(n-1)\sigma^2 - \tau^2}{m(n-m)(\sigma^2 + \tau^2)} \right\}^{\frac{1}{2}} \left(\sum_{i=1}^m Z(\mathbf{v}_i) - m\bar{Z} \right). \quad (4.24)$$

If $m = \sum_{i=1}^m \sum_{j=1}^m \mathbf{x}(\mathbf{v}_i)' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}(\mathbf{v}_j)$ in (4.23), or if $m = n$ in (4.24), then existence condition (E1) is violated, and no solution exists for the respective equations.

It can also be shown that (4.19) and (4.20) in Proposition 4.8 reduce to (4.23) and (4.24), respectively, when $B = \{s_0\}$.

2. Suppose $B = \{s_0\}$, and $A^\# = \{\mathbf{v}_1\}$ (i.e., $\{\mathbf{v}_1\}$ is the only element in the sample A that is equal, or closest, to $\{s_0\}$). Then $m \equiv |A^\#| = 1$, and (4.24) reduces to

$$\hat{S}_{oc}(s_0) = \bar{Z} + \left\{ \frac{(n-1)\sigma^2 - \tau^2}{(n-1)(\sigma^2 + \tau^2)} \right\}^{\frac{1}{2}} (Z(\mathbf{v}_1) - m\bar{Z}). \quad (4.25)$$

If, in addition, $\tau^2 = 0$, then

$$\hat{S}_{oc}(s_0) = Z(\mathbf{v}_1).$$

Now let s_0 vary over D . Consequently, for general $\tau^2 \geq 0$, ordinary constrained kriging given by (4.25) yields a piecewise constant prediction surface, constant on Voronoi polygons in \mathbb{R}^d ; if $\tau^2 = 0$, then the surface honors the data.

4.4 Constrained Kriging and Finite-population Sampling

The constrained kriging predictors (4.4) and (4.7) may be useful in contexts other than a purely spatial context. For example, we show how constrained kriging may be useful in estimating the population cumulative distribution function in a finite-population sampling context, when the data are contaminated with measurement error.

4.4.1 Finite-population Sampling

Consider the finite-population measurement-error model

$$Z(s) = S(s) + \epsilon(s); \quad s \in D, \quad (4.26)$$

where $S(\cdot)$ is a “signal” stochastic process, $\epsilon(\cdot)$ is a measurement-error process assumed to be independent of $S(\cdot)$, D is an arbitrary finite set, and $N = |D|$ is the number of *labels* or elements in the finite population.

Suppose a *sample* $A \subset D$ is drawn. The process of drawing the sample and obtaining the corresponding Z -values is called a *survey sample*. The *data* collected in a survey sample consist of both the labels and their corresponding measurements, written as a set of ordered pairs,

$$X = \{(s, Z(s)) : s \in A\}. \quad (4.27)$$

A *sampling design* (or *design*) is a probability mass function $p(\cdot)$ defined on subsets of D , such that

$$\Pr(A = a) = p(a).$$

This defines the probability that the sample a is selected.

Observe that the definition of data in (4.27) is more general than the definition given in Section 2. In this context, both A and $Z(\cdot)$ may be considered as random variables or random processes; the randomness in A depends upon the design p , and the randomness in $Z(\cdot)$ depends on some model ξ (e.g., the measurement-error spatial model (2.2)).

Model-based inference depends upon the randomness in $Z(\cdot)$ specified by the model ξ , *conditional* on the sample A (i.e., A is assumed *fixed*). For example, inference in a geostatistical context, such as is outlined in Section 2, is typically model-based since $Z(\cdot)$ is considered random and A is assumed to be fixed.

Design-based inference depends upon the randomness in the sample A , specified by the design p , *conditional* on $Z(\cdot)$ (i.e., $Z(\cdot)$ is assumed *fixed*). For example, inference in “classical” survey-sampling theory (e.g., Särndal et al., 1992) is typically design-based, where $\{Z(s) : s \in D\}$ is assumed fixed, $\{Z(s) : s \in A\}$ is assumed to be exact (i.e., measured without error), and A is random.

Measurement-error, design-based inference depends upon the randomness in the sample A specified by p and the randomness in the measurement-error process $\epsilon(\cdot)$, *conditional* on $S(\cdot)$ (i.e., $S(\cdot)$ is assumed *fixed*), where p and $\epsilon(\cdot)$ are assumed to be independent. In this modification of design-based inference, it is assumed that although the study variable is fixed, it is actually observed with error and the measurement-error component is random.

In survey-sampling theory, a lot of attention has been given to estimating the *population total* of $S(\cdot)$, defined as

$$T_S \equiv NS(D) = \sum_{s \in D} S(s),$$

where $N = |D|$. The population total T_S is linear in $S(\cdot)$ (e.g., Cochran, 1977, and Särndal et al., 1992). But, nonlinear functionals of $S(\cdot)$ such as the *population cumulative distribution function* (CDF) of $S(\cdot)$, may also be of interest. The CDF is defined as

$$F_S(t) = N^{-1} \sum_{s \in D} I(S(s) \leq t), \quad t \in \mathbb{R}, \quad (4.28)$$

where $I(\cdot)$ is the indicator function.

4.4.2 Horvitz-Thompson Estimation

In survey-sampling theory, the *Horvitz-Thompson estimator* (Horvitz and Thompson, 1952) has useful properties (e.g., Särndal et al., 1992, Sect. 2.8), such as p -unbiasedness (i.e., design-based unbiasedness).

Define the *first-order inclusion probability* of element $s \in D$ as follows:

$$\pi(s) = \Pr(s \in A) = \sum_{a: s \in a} p(a).$$

The probability that both elements $s, u \in D$ are included in a sample is given by the *second-order inclusion probability* of $s, u \in D$ as follows:

$$\pi(s, u) = \Pr(s \in A \text{ and } u \in A) = \sum_{a: s \& u \in a} p(a).$$

Note that $\pi(s, s) = \pi(s)$.

A *probability sampling design* is a design for which the following two conditions hold (Overton, 1993):

1. $\pi(s) > 0$, for all $s \in D$.
2. $\pi(s)$ is known, for all $s \in A$.

A *measurable probability sampling design* is a probability sampling design for which the following two conditions hold (Overton, 1993):

1. $\pi(s, u) > 0$, for all $s, u \in D$.
2. $\pi(s, u)$ is known, for all $s, u \in A$.

If p is a probability sampling design and $\pi(\cdot)$ is the resulting first-order inclusion probability, then the *Horvitz-Thompson estimator* (HTE) of the total T is defined as

$$\hat{T}_{S;ht} = \sum_{s \in A} \frac{S(s)}{\pi(s)} = \sum_{s \in D} \frac{S(s)}{\pi(s)} I(s \in A).$$

Its *sampling variance* is given by

$$\text{var}(\hat{T}_{S;ht}) = \sum_{s \in D} \sum_{u \in D} \frac{(\pi(s, u) - \pi(s)\pi(u))}{\pi(s)\pi(u)} S(s)S(u),$$

where recall that $\pi(\cdot, \cdot)$ is the second-order inclusion probability. If p is a measurable probability sampling design, an unbiased estimator of the sampling variance is

$$\hat{V}(\hat{T}_{S;ht}) = \sum_{s \in A} \sum_{u \in A} \frac{(\pi(s, u) - \pi(s)\pi(u))}{\pi(s, u)\pi(s)\pi(u)} S(s)S(u).$$

Consequently, if p is a probability sampling design, then the HTE of the population CDF $F_S(t)$ is

$$\hat{F}_{S;ht}(t) = \frac{1}{N} \sum_{s \in A} \frac{I(S(s) \leq t)}{\pi(s)}; \quad t \in \mathbb{R},$$

and its *sampling variance* is given by

$$\text{var}(\hat{F}_{S;ht}(t)) = \frac{1}{N^2} \sum_{s_1 \in D} \sum_{s_2 \in D} \frac{(\pi(s_1, s_2) - \pi(s_1)\pi(s_2))}{\pi(s_1)\pi(s_2)} I(S(s_1) \leq t) I(S(s_2) \leq t).$$

Furthermore, if p is a measurable probability sampling design, an unbiased estimator of the sampling variance is

$$\hat{V}(\hat{F}_{S;ht}(t)) = \frac{1}{N^2} \sum_{s \in A} \sum_{u \in A} \frac{(\pi(s, u) - \pi(s)\pi(u))}{\pi(s, u)\pi(s)\pi(u)} I(S(s) \leq t) I(S(u) \leq t).$$

4.4.3 Measurement Error and HTE Bias

Assume the finite-population measurement-error model (4.26), where the measurement-error process $\epsilon(\cdot)$ is a zero-mean white-noise process and

$$\text{var}(\epsilon(s)) \equiv \tau^2; \quad s \in D.$$

Suppose that X is a random variable measurable with respect to the joint distribution of $\{S(\cdot), \epsilon(\cdot), A\}$. Then the measurement-error, design-based expectation of X is defined as the joint conditional expectation, $E[X|S(\cdot)]$.

If $\tau^2 > 0$, then $S(\cdot)$, and hence $\hat{T}_{S;ht}$ and $\hat{F}_{S;ht}(t)$, are all unobservable. We shall now investigate whether useful inferences can be made about T_S and $F_S(t)$ from HTEs based on the process $Z(\cdot)$. Specifically, consider,

$$\hat{T}_{Z;ht} = \sum_{s \in A} \frac{Z(s)}{\pi(s)},$$

and

$$\hat{F}_{Z;ht}(t) = \frac{1}{N} \sum_{s \in A} \frac{I(Z(s) \leq t)}{\pi(s)}; \quad t \in \mathbb{R}, \quad (4.29)$$

which are HTEs, respectively, of

$$T_Z \equiv N Z(D) = \sum_{s \in D} Z(s),$$

and

$$F_Z(t) \equiv N^{-1} \sum_{s \in D} I(Z(s) \leq t); \quad t \in \mathbb{R}.$$

Suppose for the moment that $\tau^2 = 0$. Then $\epsilon(\cdot) = 0$, $Z(\cdot) = S(\cdot)$, and the joint measurement-error, design-based distribution of $\{\epsilon(\cdot), A\}$ reduces to the design p . Therefore,

$$\begin{aligned} E(\hat{T}_{Z;ht}|S(\cdot)) &= E\left[\sum_{s \in A} \frac{Z(s)}{\pi(s)} \middle| S(\cdot)\right] \\ &= \sum_{s \in D} \frac{S(s)}{\pi(s)} E[I(s \in A)] \\ &= T_S, \end{aligned}$$

and

$$\begin{aligned} E(\hat{F}_{Z;ht}(t)|S(\cdot)) &= \frac{1}{N} \sum_{s \in D} \frac{I(S(s) \leq t)}{\pi(s)} E[I(s \in A)] \\ &= F_S(t). \end{aligned}$$

Thus, if $\tau^2 = 0$, then $\hat{T}_{Z;ht}$ and $\hat{F}_{Z;ht}(t)$ are design-unbiased estimators of T_S and $F_S(t)$, respectively.

Now, suppose that $\tau^2 > 0$. Then

$$\begin{aligned} E(\hat{T}_{Z;ht}|S(\cdot)) &= E\left[E\left(\sum_{s \in A} \frac{Z(s)}{\pi(s)} \middle| S(\cdot), A\right) \middle| S(\cdot)\right] \\ &= E\left[\sum_{s \in A} \frac{S(s)}{\pi(s)} \middle| S(\cdot)\right] \\ &= T_S. \end{aligned}$$

That is, $\hat{T}_{Z;ht}$ is a measurement-error, design-unbiased estimator of T_S , and Särndal et al. (1992; Ch. 16) show that $\text{var}(\hat{T}_{Z;ht}|S(\cdot))$ can be simply partitioned into sampling-error and measurement-error components.

However, $\hat{F}_{Z;ht}(t)$ is not a measurement-error, design-unbiased estimator of $F_S(t)$, since

$$\begin{aligned} E(\hat{F}_{Z;ht}(t)|S(\cdot)) &= E[E(\hat{F}_{Z;ht}(t)|S(\cdot), \epsilon(\cdot))|S(\cdot)] \\ &= E[F_Z(t)|S(\cdot)] \\ &= \frac{1}{N} \sum_{s \in D} E(I(S(s) + \epsilon(s) \leq t)|S(\cdot)) \\ &= \frac{1}{N} \sum_{s \in D} G_\epsilon(t - S(s)) \\ &= (G_\epsilon * F_S)(t) \\ &\neq F_S(t), \text{ for } \tau^2 > 0, \end{aligned}$$

where $G_\epsilon(t) = \Pr(\epsilon(s) \leq t)$; $t \in \mathbb{R}$, and “ $*$ ” denotes convolution. If τ^2 is large, then the bias in $\hat{F}_{Z;ht}(t)$ may be substantial.

4.4.4 Alternative CDF estimators

It is only recently that the problem of the bias in CDF estimation due to measurement error has been addressed (e.g., Fuller, 1995; Stefanski and Bay, 1996). Several alternative CDF estimators with better bias properties than the HTE in (4.29) are briefly presented here.

Simulation Extrapolation Deconvolution

Stefanski and Bay (1996) use a simulation-extrapolation deconvolution argument to provide a bias-adjusted CDF estimator. They assume model (4.26) where S is fixed and $\epsilon(u_1), \dots, \epsilon(u_N)$ are i.i.d. $N(0, \tau^2)$, independent of the sampling design p . Very briefly, their method proceeds as follows: (i) add additional pseudo random errors of known variance to the original data, creating error-inflated “pseudo data” sets; (ii) recompute “pseudo CDF estimators” from the pseudo data; (iii) establish a trend in the pseudo estimators as a function of the variance of the added errors, and extrapolate backwards to the case of no measurement error.

The authors note that the theory underlying simulation-extrapolation assumes that τ^2 is “small”. Even so, they suggest this method should significantly reduce bias when τ^2 is moderate or even large.

Simplified Model

Fuller (1995) assumes the measurement-error model (4.26), where $S(u_1), \dots, S(u_N)$ are i.i.d. $N(\mu, \sigma^2)$, $\epsilon(u_1), \dots, \epsilon(u_N)$ are i.i.d. $N(0, \tau^2)$, and $\epsilon(\cdot)$ and $S(\cdot)$ are independent. By invoking this model we obtain the model CDF,

$$G_S(t) = \Phi\left(\frac{t - \mu}{\sigma}\right), \quad (4.30)$$

where $\Phi(\cdot)$ is the standard normal CDF. Suppose that a sample of size n is taken such that the labels $\{s_1, \dots, s_k\}$, with $k < n$, are selected, and n_j replicate samples are taken at label s_j ; $j = 1, \dots, k$, where $\sum_{j=1}^k n_j = n$. Define

$$\hat{\mu} \equiv \frac{1}{k} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} Z_j(s_i).$$

From the analysis of variance of the random-effects model, Fuller (1995) obtains the following variance estimator:

$$\hat{\sigma}^2 \equiv \frac{1}{r} \left(\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{Z}(s_i) - \hat{\mu})^2 - \hat{\tau}^2 \right), \quad (4.31)$$

where $\bar{Z}(s_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_j(s_i)$, $r = (k-1)^{-1}(n - n^{-1} \sum_{i=1}^k n_i^2)$, and $\hat{\tau}^2 = (\sum_{i=1}^k (n_i - 1))^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (Z_j(s_i) - \bar{Z}(s_i))^2$. It is straightforward to show that $E(\hat{\mu}) = \mu$ and $E(\hat{\sigma}^2) = \sigma^2$. The estimator in (4.31) may not be positive — in cases where $\hat{\sigma}^2$ is negative,

replace $\hat{\sigma}^2$ with a small positive number. Substituting $\hat{\mu}$ for μ and $\hat{\sigma}$ for σ in (4.30) we obtain the following CDF estimator:

$$\hat{F}_{S;sm}(t) = \Phi\left(\frac{t - \hat{\mu}}{\hat{\sigma}}\right). \quad (4.32)$$

Horvitz-Thompson Estimator with Constrained Kriging

Assume that $S(\mathbf{u}_1), \dots, S(\mathbf{u}_N)$ are i.i.d. (μ, σ^2) (i.e., each $S(\mathbf{u}_i)$ is an independent and identically distributed random variable with a mean of μ and a variance of σ^2), $\epsilon(\mathbf{u}_1), \dots, \epsilon(\mathbf{u}_N)$ are i.i.d. $(0, \tau^2)$, and $\epsilon(\cdot)$ and $S(\cdot)$ are independent. Since $S(\cdot)$ is a second-order-stationary uncorrelated stochastic process, we can write the ordinary kriging constrained predictor (see Section 4.3) of $S(\mathbf{s})$; $\mathbf{s} \in A$, as

$$\hat{S}_{oc}(\mathbf{s}) = \bar{Z} + \left\{ \frac{(n-1)\sigma^2 - \tau^2}{(n-1)(\sigma^2 + \tau^2)} \right\}^{\frac{1}{2}} (Z(\mathbf{s}) - \bar{Z}); \quad \mathbf{s} \in A,$$

where $A = \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset D$ is the sample, and $\bar{Z} = n^{-1} \sum_{i=1}^n Z(\mathbf{s}_i)$ is the BLUE of μ .

By definition of the ordinary constrained kriging predictor, we have

$$E[\hat{S}_{oc}(\mathbf{s})] = \mu \quad \text{and} \quad \text{var}(\hat{S}_{oc}(\mathbf{s})) = \sigma^2; \quad \mathbf{s} \in A.$$

However, $\hat{S}_{oc}(\cdot) \equiv \{\hat{S}_{oc}(\mathbf{s}) : \mathbf{s} \in A\}$ is *not* an uncorrelated process, since

$$\begin{aligned} \text{cov}(\hat{S}_{oc}(\mathbf{s}), \hat{S}_{oc}(\mathbf{u})) &= \frac{\sigma^2 + \tau^2}{n} + \left(\frac{(n-1)\sigma^2 - \tau^2}{(n-1)(\sigma^2 + \tau^2)} \right) \left(-\frac{\sigma^2 + \tau^2}{n} \right); \quad \mathbf{s} \neq \mathbf{u} \\ &= \frac{\tau^2}{n-1}; \quad \mathbf{s} \neq \mathbf{u}. \end{aligned} \quad (4.33)$$

If $S(\cdot)$ and $\epsilon(\cdot)$ are Gaussian processes, then $\hat{S}_{oc}(\mathbf{s}) \stackrel{d}{=} S(\mathbf{s})$; $\mathbf{s} \in A$, although (4.33) shows that the joint distributions of $(\hat{S}_{oc}(\mathbf{s}_1), \dots, \hat{S}_{oc}(\mathbf{s}_n))$ and $(S(\mathbf{s}_1), \dots, S(\mathbf{s}_n))$ are not equal. However, if $\tau^2/(n-1)$ is small, we may be justified in constructing a Horvitz-Thompson estimator of the CDF, using the measurement-error-adjusted random variables $(\hat{S}_{oc}(\mathbf{s}_1), \dots, \hat{S}_{oc}(\mathbf{s}_n))$ in place of $(Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))$ in (4.29). Hence we define the HTE CDF estimator, using constrained kriging, as

$$\hat{F}_{S;hc}(t) \equiv \frac{1}{N} \sum_{\mathbf{s} \in A} \frac{I(\hat{S}_{oc}(\mathbf{s}) \leq t)}{\pi(\mathbf{s})}; \quad t \in \mathbb{R}.$$

Proposition 4.10 *If $S(\mathbf{u}_1), \dots, S(\mathbf{u}_N)$ are i.i.d. $N(\mu, \sigma^2)$ and $\epsilon(\mathbf{u}_1), \dots, \epsilon(\mathbf{u}_N)$ are i.i.d. $N(0, \tau^2)$, then*

$$E[\hat{F}_{S;hc}(t)|A] = E[\hat{F}_{S;ht}(t)|A].$$

Furthermore, because $(\hat{S}_{oc}(\mathbf{s}), \hat{S}_{oc}(\mathbf{u}))$ is bivariate normal for any $\mathbf{s} \neq \mathbf{u}$, then

$$\begin{aligned} \text{var}(\hat{F}_{S;hc}(t)|A) - \text{var}(\hat{F}_{S;ht}(t)|A) = \\ \frac{1}{N^2} \left[\Phi_2\left(\frac{t-\mu}{\sigma}, \frac{t-\mu}{\sigma}; \rho\right) - \left\{ \Phi\left(\frac{t-\mu}{\sigma}\right) \right\}^2 \right] \sum_{\mathbf{s} \in A} \sum_{\substack{\mathbf{u} \in A \\ \mathbf{u} \neq \mathbf{s}}} \frac{1}{\pi(\mathbf{s})\pi(\mathbf{u})}, \end{aligned}$$

where $\hat{F}_{S;ht}(t) = \frac{1}{N} \sum_{\mathbf{s} \in A} \frac{S(\mathbf{s})}{\pi(\mathbf{s})}$ (which is unobservable if $\tau^2 > 0$), $\Phi(\cdot)$ is the standard normal CDF, $\Phi_2(\cdot, \cdot; \rho)$ is the CDF of a bivariate normal distribution with zero means, unit variances, and correlation $\rho \equiv \text{corr}(\hat{S}_{oc}(\mathbf{s}), \hat{S}_{oc}(\mathbf{u})) = \{\tau^2/(\sigma^2(n-1))\}^{\frac{1}{2}}$.

Proof: If $S(\mathbf{u}_1), \dots, S(\mathbf{u}_N)$ are i.i.d. $N(\mu, \sigma^2)$ and $\epsilon(\mathbf{u}_1), \dots, \epsilon(\mathbf{u}_N)$ are i.i.d. $N(0, \tau^2)$, then $\hat{S}_{oc}(\mathbf{s})$ is distributed as $N(\mu, \sigma^2)$, for all $\mathbf{s} \in A$. Therefore,

$$\begin{aligned} E[I(\hat{S}_{oc}(\mathbf{s}) \leq t)] &= \Phi\left(\frac{t-\mu}{\sigma}\right); \quad \mathbf{s} \in A \\ &= E[I(S(\mathbf{s}) \leq t)]; \quad \mathbf{s} \in A, \end{aligned}$$

and hence

$$E[\hat{F}_{S;hc}(t)|A] = E[\hat{F}_{S;ht}(t)|A].$$

Now,

$$\begin{aligned} \text{var}[I(\hat{S}_{oc}(\mathbf{s}) \leq t)] &= \Phi\left(\frac{t-\mu}{\sigma}\right) \left[1 - \Phi\left(\frac{t-\mu}{\sigma}\right) \right]; \quad \mathbf{s} \in A \\ &= \text{var}[I(S(\mathbf{s}) \leq t)]; \quad \mathbf{s} \in A. \end{aligned}$$

By assumption, all random variables are multivariate normal and hence the pair of linear combinations, $(\hat{S}_{oc}(\mathbf{s}), \hat{S}_{oc}(\mathbf{u}))$ is bivariate normal for any $\mathbf{s} \neq \mathbf{u}$. Then

$$\begin{aligned} \text{cov}[I(\hat{S}_{oc}(\mathbf{s}) \leq t), I(\hat{S}_{oc}(\mathbf{u}) \leq t)] &= \Pr(\hat{S}_{oc}(\mathbf{s}) \leq t, \hat{S}_{oc}(\mathbf{u}) \leq t) - \Pr(\hat{S}_{oc}(\mathbf{s}) \leq t) \Pr(\hat{S}_{oc}(\mathbf{u}) \leq t) \\ &= \left[\Phi_2\left(\frac{t-\mu}{\sigma}, \frac{t-\mu}{\sigma}; \rho\right) - \left\{ \Phi\left(\frac{t-\mu}{\sigma}\right) \right\}^2 \right], \end{aligned}$$

where $\rho = \text{corr}(\hat{S}_{oc}(s), \hat{S}_{oc}(u)) = \{\tau^2/(\sigma^2(n-1))\}^{\frac{1}{2}}$; $s \neq u$. Therefore,

$$\begin{aligned} \text{var}[\hat{F}_{S;hc}(t)|A] &= \frac{1}{N^2} \left\{ \Phi\left(\frac{t-\mu}{\sigma}\right) \left[1 - \Phi\left(\frac{t-\mu}{\sigma}\right)\right] \right\} \sum_{s \in A} \left(\frac{1}{\pi(s)}\right)^2 \\ &\quad + \frac{1}{N^2} \left[\Phi_2\left(\frac{t-\mu}{\sigma}, \frac{t-\mu}{\sigma}; \rho\right) - \left\{ \Phi\left(\frac{t-\mu}{\sigma}\right) \right\}^2 \right] \sum_{s \in A} \sum_{\substack{u \in A \\ u \neq s}} \frac{1}{\pi(s)\pi(u)}. \end{aligned}$$

Now, $S(u_1), \dots, S(u_N)$ i.i.d. $N(\mu, \sigma^2)$ implies that

$$\Pr(S(s) \leq t, S(u) \leq t) = \Pr(S(s) \leq t) \Pr(S(u) \leq t); \quad s \neq u,$$

and hence $\text{cov}[I(S(s) \leq t), I(S(u) \leq t)] = 0$; $s \neq u$. Thus,

$$\text{var}[\hat{F}_{S;ht}(t)|A] = \frac{1}{N^2} \left\{ \Phi\left(\frac{t-\mu}{\sigma}\right) \left[1 - \Phi\left(\frac{t-\mu}{\sigma}\right)\right] \right\} \sum_{s \in A} \left(\frac{1}{\pi(s)}\right)^2,$$

and the result follows. \square

Some consequences of Proposition 4.10 follow:

1. If $\rho \equiv \{\tau^2/(\sigma^2(n-1))\}^{\frac{1}{2}} \simeq 0$ then $\text{var}[\hat{F}_{S;hc}(t)|A] \simeq \text{var}[\hat{F}_{S;ht}(t)|A]$.
2. if p is a non-replacement, equiprobable probability sampling design of fixed size n , then $\pi(s) \equiv n/N$. Therefore,

$$E[\hat{F}_{S;hc}(t)|A] = E[\hat{F}_{S;ht}(t)|A] = \Phi\left(\frac{t-\mu}{\sigma}\right),$$

and

$$\text{var}(\hat{F}_{S;hc}(t)|A) - \text{var}(\hat{F}_{S;ht}(t)|A) = \frac{n-1}{n} \left[\Phi_2\left(\frac{t-\mu}{\sigma}, \frac{t-\mu}{\sigma}; \rho\right) - \left\{ \Phi\left(\frac{t-\mu}{\sigma}\right) \right\}^2 \right].$$

Typically, σ^2 and τ^2 are unknown. Assuming replicate samples at $k < n$ sites (i.e., s_1, \dots, s_k), we may obtain $\hat{\sigma}^2$ and $\hat{\tau}^2$ in the same manner as described in the vicinity of (4.31). Following the usual practice in geostatistics of estimating the covariance parameters and “plugging” them in to predictors (see Section 3.2.2), we obtain the ordinary constrained kriging predictor (see Section 4.3), as follows,

$$\hat{S}_{oc}(s_j) = \bar{Z}_{..} + \left\{ \frac{(n-1)\hat{\sigma}^2 - \hat{\tau}^2}{n_j(n-n_j)(\hat{\sigma}^2 + \hat{\tau}^2)} \right\}^{\frac{1}{2}} \left(\sum_{i=1}^{n_j} Z_i(s_j) - n_j \bar{Z}_{..} \right); \quad j = 1, \dots, k,$$

where $\bar{Z}_{..} = \frac{1}{k} \sum_{j=1}^k \frac{1}{n_j} \sum_{i=1}^{n_j} Z_i(\mathbf{s}_j)$, and $\hat{\sigma}^2$ and $\hat{\tau}^2$ are defined as in (4.31). If the n sampling units are chosen according to a (with-replacement) probability sampling design, then we define the HTE CDF estimator, using constrained kriging, as

$$\hat{F}_{S;hc}(t) \equiv \frac{1}{N} \sum_{j=1}^k n_j \frac{I(\hat{S}_{oc}(\mathbf{s}_j) \leq t)}{\pi(\mathbf{s}_j)}; \quad t \in \mathbb{R}.$$

4.5 Constrained Kriging and Conditional Simulation

Simple kriging or ordinary kriging predictors are generally smoother than their corresponding predictands. Recognizing this, Journel (1974) constructed a simulation algorithm that yielded more realistic realizations of the spatial process of interest (in his case, this was $Z(\cdot)$) than the surface defined by simple kriging or ordinary kriging. It is called *conditional simulation* because all realizations are conditioned to go through the data \mathbf{Z} . In his original formulation, conditional simulation was based on simple kriging, but another version based on ordinary kriging was formulated by Journel and Huijbregts (1978, Ch. VII). In either formulation, the conditional-simulation process honors the data, although in the measurement-error version of these processes discussed below, this is not necessarily the case. The construction of these conditional-simulation processes depends upon the orthogonality properties of simple kriging and ordinary kriging.

4.5.1 Orthogonality Properties of Simple Kriging and Ordinary Kriging

Consider the measurement-error spatial model (2.2). For ease of presentation, for this entire section dealing with conditional simulation, we will assume that $S(\cdot)$ (and hence $Z(\cdot)$) is second-order stationary with $E(S(\mathbf{s})) \equiv \mu$, $\text{var}(S(\mathbf{s})) \equiv \sigma^2$, and $\text{var}(\epsilon(\mathbf{s})) \equiv \tau^2$. In fact, all that is needed for these simulation algorithms is knowledge of the mean function and the covariance function, whether they be stationary or not.

Consider the simple kriging predictor of $S(\mathbf{s})$ given by

$$\hat{S}_{sk}(\mathbf{s}) \equiv \mu + \mathbf{c}(\mathbf{s})' \Sigma^{-1} (\mathbf{Z} - \mu \mathbf{1}); \quad \mathbf{s} \in D,$$

and the ordinary kriging predictor of $S(s)$ given by

$$\hat{S}_{ok}(s) \equiv \hat{\mu} + c(s)' \Sigma^{-1} (Z - \hat{\mu} \mathbf{1}); \quad s \in D,$$

where $c(s) = (C(s_1, s), \dots, C(s_n, s))'$, $\Sigma = \text{var}(Z)$, and $\hat{\mu} = (\mathbf{1}' \Sigma^{-1} Z) / (\mathbf{1}' \Sigma^{-1} \mathbf{1})$.

Let A be an $n \times m$ matrix. Then,

$$\begin{aligned} \text{cov}(S(s) - \hat{S}_{sk}(s), A'Z) &= c(s)' A - c(s)' \Sigma^{-1} \Sigma A \\ &= 0; \quad s \in D, \end{aligned}$$

for all A . That is, the simple kriging prediction error $(S(s) - \hat{S}_{sk}(s))$ is orthogonal to *any* linear function of Z . As a consequence,

$$\begin{aligned} \text{cov}[(S(s) - \hat{S}_{sk}(s)), (S(u) - \hat{S}_{sk}(u))] &= \text{cov}[(S(s) - \hat{S}_{sk}(s)), S(u)] \\ &= \text{cov}(S(s), S(u)) - \text{cov}(\hat{S}_{sk}(s), \hat{S}_{sk}(u)), \quad (4.34) \end{aligned}$$

since $\text{cov}(\hat{S}_{sk}(s), S(u)) = c(s)' \Sigma^{-1} c(u) = \text{cov}(\hat{S}_{sk}(s), \hat{S}_{sk}(u)); \quad s, u \in D$.

In the case of ordinary kriging and for matrices A such that $A' \mathbf{1} = 0$,

$$\begin{aligned} \text{cov}(S(s) - \hat{S}_{ok}(s), A'Z) &= c(s)' A - \mathbf{1}' A / (\mathbf{1}' \Sigma^{-1} \mathbf{1}) + \mathbf{1}' A (c(s)' \Sigma^{-1} \mathbf{1}) / (\mathbf{1}' \Sigma^{-1} \mathbf{1}) \\ &\quad - c(s)' \Sigma^{-1} \Sigma A \\ &= \mathbf{1}' A (c(s)' \Sigma^{-1} \mathbf{1} - \mathbf{1}) / (\mathbf{1}' \Sigma^{-1} \mathbf{1}) \\ &= 0; \quad s \in D. \end{aligned}$$

That is, the ordinary kriging prediction error $(S(s) - \hat{S}_{ok}(s))$ is orthogonal to *any contrast* among the elements of Z .

4.5.2 Conditional Simulation and Measurement Error

Journel (1974) used the orthogonality properties of simple kriging to construct a conditional-simulation algorithm for $Z(\cdot)$ based on simple kriging, and Journel and Huijbregts (1978, Ch. VII) used the orthogonality properties of ordinary kriging to construct a conditional simulation algorithm for $Z(\cdot)$ based on ordinary kriging. We modify their methods to construct conditional-simulation algorithms for the signal $S(\cdot)$ based on simple kriging and on ordinary kriging, respectively.

Suppose that we can generate (nonconditional) simulation processes $S_N(\cdot)$ and $\epsilon_N(\cdot)$, independent of one another, such that

$$E(S_N(s)) = E(S(s)) = \mu; \quad s \in D,$$

$$E(\epsilon_N(s)) = E(\epsilon(s)) = 0; \quad s \in D,$$

and

$$\text{cov}(S_N(s), S_N(u)) = \text{cov}(S(s), S(u)); \quad s, u \in D,$$

$$\text{cov}(\epsilon_N(s), \epsilon_N(u)) = \text{cov}(\epsilon(s), \epsilon(u)); \quad s, u \in D.$$

That is, the expectations, variances, and covariances of $S_N(\cdot)$ match with those of $S(\cdot)$, and likewise for $\epsilon_N(\cdot)$ with $\epsilon(\cdot)$. Furthermore, they are generated in such a way that they are independent of $S(\cdot)$ and $\epsilon(\cdot)$.

Now, the distribution that really is of interest is the posterior distribution of $S(\cdot)$ (i.e., $S(\cdot)|\mathbf{Z}$), since under any loss function the best predictor depends upon it. For example, under squared-error loss, the best predictor of $S(s)$ is $E(S(s)|\mathbf{Z})$. What we require is a simulation process whose distribution at least approximately matches that of $S(\cdot)|\mathbf{Z}$. In this respect, the nonconditionally simulated processes do not work since the independence of $(S(\cdot), \epsilon(\cdot))$ and $S_N(\cdot)$ implies that

$$\begin{aligned} E(S_N(s)|\mathbf{Z}) &= E(S_N(s)) \\ &= E(S(s)) \\ &\neq E(S(s)|\mathbf{Z}). \end{aligned}$$

Conditional Simulation Based On Simple Kriging

Define the conditional simulation process of $S(\cdot)$, based on simple kriging, as

$$S_{C,sk}(s) \equiv \hat{S}_{sk}(s) + (S_N(s) - \hat{S}_{N,sk}(s)); \quad s \in D, \quad (4.35)$$

$$= S_N(s) + \mathbf{c}(s)' \Sigma^{-1}(\mathbf{Z} - \mathbf{Z}_N); \quad s \in D, \quad (4.36)$$

since $\hat{S}_{sk}(s) = \mu + \mathbf{c}(s)' \Sigma^{-1}(\mathbf{Z} - \mu \mathbf{1})$, and $\hat{S}_{N,sk}(s) \equiv \mu + \mathbf{c}(s)' \Sigma^{-1}(\mathbf{Z}_N - \mu \mathbf{1})$, where $\mathbf{Z}_N = (Z_N(s_1), \dots, Z_N(s_n))'$, and $Z_N(s) \equiv S_N(s) + \epsilon_N(s)$; $s \in D$. Observe that $S_{C,sk}(s)$ depends on

μ only through $S_N(s)$ and Z_N (i.e., knowledge of μ is required to generate $S_N(\cdot)$). Now, (4.36) implies that

$$E(S_{C,sk}(s)) = E(S_N(s)) = E(S(s)); \quad s \in D,$$

and, the orthogonality property (4.34) implies that

$$\begin{aligned} \text{cov}(S_{C,sk}(s), S_{C,sk}(u)) &= \text{cov}(\hat{S}_{sk}(s), \hat{S}_{sk}(u)) + \text{cov}[(S_N(s) - \hat{S}_{N,sk}(s)), (S_N(u) - \hat{S}_{N,sk}(u))] \\ &= \text{cov}(\hat{S}_{sk}(s), \hat{S}_{sk}(u)) + \text{cov}(S_N(s), (S_N(u) - \hat{S}_{N,sk}(u))) - \text{cov}(\hat{S}_{N,sk}(s), \hat{S}_{N,sk}(u)) \\ &= \text{cov}(S_N(s), (S_N(u) - \hat{S}_{N,sk}(u))) \\ &= \text{cov}(S(s), (S(u) - \hat{S}(u))); \quad s, u \in D. \end{aligned}$$

That is, the means and covariance functions of $S_{C,sk}(\cdot)$ and $S(\cdot)$ match.

Further,

$$\begin{aligned} E(S_{C,sk}(s)|Z) &= E(\hat{S}_{sk}(s)|Z) + E[(S_N(s) - \hat{S}_{N,sk}(s))|Z] \\ &= \hat{S}_{sk}(s) + E(S_N(s) - \hat{S}_{N,sk}(s)) \\ &= \hat{S}_{sk}(s); \quad s \in D, \end{aligned} \tag{4.37}$$

and

$$\begin{aligned} \text{var}(S_{C,sk}(s)|Z) &= \text{var}(\hat{S}_{sk}(s)|Z) + \text{var}[(S_N(s) - \hat{S}_{N,sk}(s))|Z] \\ &= 0 + \text{var}(S_N(s) - \hat{S}_{N,sk}(s)) \\ &= \text{var}(S(s) - \hat{S}(s)) \\ &\equiv \text{MSPE}(\hat{S}_{sk}(s)); \quad s \in D. \end{aligned} \tag{4.38}$$

Thus, if $(S_N(\cdot), \epsilon_N(\cdot))$ and $(S(\cdot), \epsilon(\cdot))$ are both bivariate Gaussian processes then, from (4.37) and (3.9),

$$E(S_{C,sk}(s)|Z) = \hat{S}_{sk}(s) = E(S(s)|Z); \quad s \in D,$$

and, from (4.38) and (3.8),

$$\begin{aligned} \text{var}(S_{C,sk}(s)|Z) &= \text{MSPE}(\hat{S}_{sk}(s)) \\ &= \sigma^2 - \mathbf{c}(s)' \Sigma^{-1} \mathbf{c}(s) \\ &= \text{var}(S(s)|Z); \quad s \in D. \end{aligned}$$

That is, if $(S_N(\cdot), \epsilon_N(\cdot))$ and $(S(\cdot), \epsilon(\cdot))$ are both bivariate Gaussian processes, then the conditional means and conditional variances also match.

The MSPE of $S_{C,sk}(s)$ is given by

$$\begin{aligned}
 \text{MSPE}(S_{C,sk}(s)) &\equiv E[(S(s) - S_{C,sk}(s))^2] \\
 &= \text{var}(S(s) - S_{C,sk}(s)) \\
 &= 2(\sigma^2 - \text{cov}(S(s), S_{C,sk}(s))) \\
 &= 2(\sigma^2 - \text{cov}(S(s), \hat{S}_{sk}(s))) \\
 &= 2(\sigma^2 - \mathbf{c}(s)' \Sigma^{-1} \mathbf{c}(s)) \\
 &= 2\text{MSPE}(\hat{S}_{sk}(s)); \quad s \in D,
 \end{aligned}$$

which suggests that for the purpose of *prediction*, $\hat{S}_{sk}(s)$ is to be preferred to $S_{C,sk}(s)$.

If $\tau^2 = 0$, then for any $\mathbf{s}_i \in A \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ it is easy to show that simple kriging honors the data (the proof can easily be adapted from Proposition 3.1, which shows that ordinary kriging honors the data when $\tau^2 = 0$). That is, if $\tau^2 = 0$ and hence $\epsilon(\cdot) = 0$, then

$$\begin{aligned}
 S_{C,sk}(\mathbf{s}_i) &= \hat{S}_{sk}(\mathbf{s}_i) + (S_N(\mathbf{s}_i) - \hat{S}_{N,sk}(\mathbf{s}_i)) \\
 &= S(\mathbf{s}_i); \quad \mathbf{s}_i \in A \\
 &= Z(\mathbf{s}_i); \quad \mathbf{s}_i \in A.
 \end{aligned}$$

Conditional Simulation Based On Ordinary Kriging

Define the conditional-simulation process of $S(\cdot)$ based on ordinary kriging as

$$S_{C,ok}(s) \equiv \hat{S}_{ok}(s) + (S_N(s) - \hat{S}_{N,ok}(s)); \quad s \in D, \quad (4.39)$$

$$\begin{aligned}
 &= \hat{\mu} + \mathbf{c}(s)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu} \mathbf{1}) + S_N(s) - \hat{\mu}_N - \mathbf{c}(s)' \Sigma^{-1} (\mathbf{Z}_N - \hat{\mu}_N \mathbf{1}) \\
 &= S_N(s) + \mathbf{c}(s)' \Sigma^{-1} (\mathbf{Z} - \mathbf{Z}_N) + (1 - \mathbf{c}(s)' \Sigma^{-1} \mathbf{1})(\hat{\mu} - \hat{\mu}_N) \\
 &= S_{C,sk}(s) + (1 - \mathbf{c}(s)' \Sigma^{-1} \mathbf{1})(\hat{\mu} - \hat{\mu}_N); \quad s \in D, \quad (4.40)
 \end{aligned}$$

since $\hat{S}_{ok}(s) = \hat{\mu} + \mathbf{c}(s)' \Sigma^{-1} (\mathbf{Z} - \hat{\mu} \mathbf{1})$ and $\hat{S}_{N,ok}(s) = \hat{\mu}_N + \mathbf{c}(s)' \Sigma^{-1} (\mathbf{Z}_N - \hat{\mu}_N \mathbf{1})$, where $\hat{\mu}_N = (\mathbf{1}' \Sigma^{-1} \mathbf{Z}_N) / (\mathbf{1}' \Sigma^{-1} \mathbf{1})$; $s \in D$. Now,

$$E(S_{C,ok}(s)) = E(S_N(s)) = E(S(s)); \quad s \in D,$$

but, from (4.40),

$$\begin{aligned}\text{cov}(S_{C,ok}(s), S_{C,ok}(u)) &\neq \text{cov}(S_{C,sk}(s), S_{C,sk}(u)) \\ &= \text{cov}(S(s), (S(u))); \quad s, u \in D.\end{aligned}\quad (4.41)$$

However,

$$\begin{aligned}\text{var}(S_{C,ok}(s) - S_{C,ok}(u)) &= \text{var}(S_{C,sk}(s) - S_{C,sk}(u)) \\ &= \text{var}(S(s)) + \text{var}(S(u)) - 2\text{cov}(S(s), S(u)) \\ &= \text{var}(S(s) - (S(u))); \quad s, u \in D.\end{aligned}$$

That is, the expectations and variograms of $S_{C,ok}(\cdot)$ and $S(\cdot)$ match, but their covariance functions do not. Further,

$$\begin{aligned}E(S_{C,ok}(s)|Z) &= E(S_{C,sk}(s)|Z) + E[(1 - c(s)' \Sigma^{-1} \mathbf{1})(\hat{\mu} - \hat{\mu}_N)|Z] \\ &= E(S_{C,sk}(s)|Z) + (1 - c(s)' \Sigma^{-1} \mathbf{1})(\hat{\mu} - \mu) \\ &\neq E(S_{C,sk}(s)|Z); \quad s \in D,\end{aligned}\quad (4.42)$$

unless $c(s)' \Sigma^{-1} \mathbf{1} = 1$ or $\hat{\mu} = \mu$, and

$$\begin{aligned}\text{var}(S_{C,ok}(s)|Z) &= 0 + \text{var}(S_N(s) - \hat{S}_{N,ok}(s)) \\ &= \text{MSPE}(\hat{S}_{N,ok}(s)) \\ &= \text{MSPE}(\hat{S}_{ok}(s)) \\ &= \text{MSPE}(\hat{S}_{sk}(s)) + (1 - c(s)' \Sigma^{-1} \mathbf{1})^2 / (\mathbf{1}' \Sigma^{-1} \mathbf{1}) \\ &\geq \text{MSPE}(\hat{S}_{sk}(s)); \quad s \in D.\end{aligned}\quad (4.43)$$

Therefore, even if $(S_N(\cdot), \epsilon_N(\cdot))$ and $(S(\cdot), \epsilon(\cdot))$ are both bivariate Gaussian processes, $S_{C,ok}(\cdot)|Z$ is distributed differently from $S(\cdot)|Z$.

Since $(S(\cdot), \epsilon(\cdot))$ and $(S_N(\cdot), \epsilon_N(\cdot))$ are independent, the MSPE of $S_{C,ok}(s)$ is given by

$$\begin{aligned}\text{MSPE}(S_{C,ok}(s)) &\equiv E[(S(s) - S_{C,ok}(s))^2] \\ &= \text{var}(S(s) - S_{C,ok}(s))\end{aligned}$$

$$\begin{aligned}
&= \text{var}(S(\mathbf{s}) - \hat{S}_{ok}(\mathbf{s})) + \text{var}(S_N(\mathbf{s}) - \hat{S}_{N,ok}(\mathbf{s})) \\
&= 2\text{var}(S(\mathbf{s}) - \hat{S}_{ok}(\mathbf{s})) \\
&= 2\text{MSPE}(\hat{S}_{ok}(\mathbf{s})) \\
&= 2(\text{MSPE}(\hat{S}_{sk}(\mathbf{s})) + (1 - \mathbf{c}(\mathbf{s})'\Sigma^{-1}\mathbf{1})^2/(\mathbf{1}'\Sigma^{-1}\mathbf{1})) \\
&\geq 2\text{MSPE}(\hat{S}_{sk}(\mathbf{s})) \\
&= \text{MSPE}(S_{C,sk}(\mathbf{s})); \quad \mathbf{s} \in D.
\end{aligned} \tag{4.44}$$

Now, since both $S_{C,sk}(\mathbf{s})$ and $S_{C,ok}(\mathbf{s})$ depend on μ in exactly the same way (i.e., through $S_N(\cdot)$ and \mathbf{Z}_N), results (4.41), (4.42), (4.43), and (4.44) suggest that conditional simulation based on simple kriging is *always* to be preferred to conditional simulation based on ordinary kriging.

4.5.3 Conditional Simulation and Constrained Kriging

Consider a conditional simulation of $S(B)$ based on simple kriging, given by

$$S_{C,sk}(B) = S_N(B) + \mathbf{c}(B)'\Sigma^{-1}(\mathbf{Z} - \mathbf{Z}_N); \quad B \subset D, \tag{4.45}$$

where $S_N(B) \equiv |B|^{-1} \int_B S_N(\mathbf{s}) d\mathbf{s}$, and $|B| = \int_B 1 d\mathbf{s} > 0$; and consider the ordinary constrained kriging predictor of $S(B)$, given by

$$\hat{S}_{oc}(B) \equiv \hat{\mu} + k\mathbf{c}(B)'\Sigma^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1}); \quad B \subset D.$$

The conditional simulation (4.45) shares some properties with constrained kriging, namely, its mean and variance match those of $S(B)$. Therefore, $g(S_{C,sk}(B))$ is approximately unbiased for $g(S(B))$ if g is smooth, and exactly unbiased if $(S_N(\cdot), \epsilon_N(\cdot))$ and $(S(\cdot), \epsilon(\cdot))$ are bivariate Gaussian processes. This begs the question: Do conditions exist for which $\text{MSPE}(S_{C,sk}(B)) < \text{MSPE}(\hat{S}_{oc}(B))$?

Proposition 4.11 *Suppose that $S(\cdot)$ is first-order stationary with $E(S(B)) \equiv \mu$ and $\text{var}(S(B)) \equiv \sigma_B^2$, and suppose that the ordinary constrained kriging predictor $\hat{S}_{oc}(B)$ of $S(B)$ exists. Then,*

$$\text{MSPE}(S_{C,sk}(B)) < \text{MSPE}(\hat{S}_{oc}(B)) \tag{4.46}$$

if and only if

$$\begin{aligned} & \left| (\mathbf{c}(B)' \Sigma^{-1} \mathbf{1}) \sigma_B^2 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \right| > \\ & \left\{ \left(\sigma_B^2 - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \right) \left(\sigma_B^2 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \right) (\mathbf{1}' \Sigma^{-1} \mathbf{1}) \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \right\}^{\frac{1}{2}}, \end{aligned} \quad (4.47)$$

provided that

$$(\mathbf{1}' \Sigma^{-1} \mathbf{1}) (\mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B)) > \mathbf{c}(B)' \Sigma^{-1} \mathbf{1}. \quad (4.48)$$

Proof: Define

$$\begin{aligned} \alpha & \equiv (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} / \sigma_B^2 = \text{var}(\hat{\mu}) / \text{var}(S(B)), \\ \beta & \equiv \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) / \sigma_B^2 = \text{var}(\hat{S}_{sk}(B)) / \text{var}(S(B)), \\ \gamma & \equiv \mathbf{c}(B)' \Sigma^{-1} \mathbf{1} = \text{cov}(S(B), \hat{\mu}) / \text{var}(\hat{\mu}). \end{aligned}$$

Observe that the existence of $\hat{S}_{oc}(B)$ implies that $\text{var}(S(B)) \geq \text{var}(\hat{\mu})$, that is, $0 < \alpha \leq 1$. Also observe that $0 < \beta \leq 1$, since $0 \leq \text{MSPE}(\hat{S}_{sk}(B)) = \text{var}(S(B)) - \text{var}(\hat{S}_{sk}(B))$ implies that $\text{var}(\hat{S}_{sk}(B)) \leq \text{var}(S(B))$.

Now,

$$\text{MSPE}(S_{C,sk}(B)) = 2(\sigma_B^2 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B)) = 2\sigma_B^2(1 - \beta),$$

and

$$\begin{aligned} \text{MSPE}(\hat{S}_{oc}(B)) &= 2 \left(\sigma_B^2 - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \mathbf{c}(B)' \Sigma^{-1} \mathbf{1} \right) \\ &\quad - 2 \left\{ \left(\sigma_B^2 - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \right) \left(\mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) - (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} (\mathbf{c}(B)' \Sigma^{-1} \mathbf{1})^2 \right) \right\}^{\frac{1}{2}} \\ &= 2\sigma_B^2 \left(1 - \alpha\gamma - \left\{ (1 - \alpha)(\beta - \alpha\gamma^2) \right\}^{\frac{1}{2}} \right). \end{aligned}$$

Therefore,

$$\text{MSPE}(S_{C,sk}(B)) - \text{MSPE}(\hat{S}_{oc}(B)) = 2\sigma_B^2 \left(\alpha\gamma - \beta + \left\{ (1 - \alpha)(\beta - \alpha\gamma^2) \right\}^{\frac{1}{2}} \right) < 0$$

if and only if

$$\beta - \alpha\gamma > \left\{ (1 - \alpha)(\beta - \alpha\gamma^2) \right\}^{\frac{1}{2}}. \quad (4.49)$$

Note that the right-hand side of (4.49) is well defined, since $0 < \alpha \leq 1$ and $\gamma^2 \leq \beta/\alpha$ by the Cauchy-Schwarz inequality. Now, inequality (4.49) may be satisfied only if

$$\beta - \alpha\gamma > 0, \quad (4.50)$$

where (4.50) is identical to (4.48).

If $\beta - \alpha\gamma > 0$, then (4.49) holds if and only if

$$(\beta - \alpha\gamma)^2 > (1 - \alpha)(\beta - \alpha\gamma^2),$$

or, equivalently,

$$f(\gamma) \equiv \alpha\gamma^2 - 2\alpha\beta\gamma - \beta(1 - \alpha - \beta) > 0,$$

or, equivalently, γ lies outside the interval with the endpoints,

$$\beta \pm \{\beta(1 - \beta)(1 - \alpha)/\alpha\}^{\frac{1}{2}}, \quad (4.51)$$

or, equivalently, $\mathbf{c}(B)' \Sigma^{-1} \mathbf{1}$ lies outside the interval with endpoints,

$$\begin{aligned} & \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) / \sigma_B^2 \\ & \pm \left\{ (1' \Sigma^{-1} \mathbf{1}) (1 - (1' \Sigma^{-1} \mathbf{1})^{-1} / \sigma_B^2) \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) (1 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) / \sigma_B^2) \right\}^{\frac{1}{2}} \\ & = \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) / \sigma_B^2 \\ & \pm \left\{ \left(\sigma_B^2 - (1' \Sigma^{-1} \mathbf{1})^{-1} \right) \left(\sigma_B^2 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \right) (1' \Sigma^{-1} \mathbf{1}) \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \right\}^{\frac{1}{2}} / \sigma_B^2, \end{aligned}$$

giving the result (4.47). \square

What are the implications of Proposition 4.11? Assume that $\hat{S}_{oc}(B)$ exists (i.e., this implies that $\sigma_B^2 \equiv \text{var}(S(B)) \geq \text{var}(\hat{\mu})$). And, from (3.8) we have,

$$\begin{aligned} \text{MSPE}(\hat{S}_{sk}(B)) &= \sigma_B^2 - \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B) \\ &= \sigma_B^2 - \text{var}(\hat{S}_{sk}(B)) \\ &\geq 0, \end{aligned}$$

which implies that $\sigma_B^2 \geq \text{var}(\hat{S}_{sk}(B))$. Therefore,

$$\begin{aligned} \sigma_B^2 &\geq \max\{\text{var}(\hat{\mu}), \text{var}(\hat{S}_{sk}(B))\} \\ &= \max\{(1' \Sigma^{-1} \mathbf{1})^{-1}, \mathbf{c}(B)' \Sigma^{-1} \mathbf{c}(B)\}. \end{aligned}$$

Observe that as σ_B^2 decreases towards $\max\{(1'\Sigma^{-1}1)^{-1}, \mathbf{c}(B)'\Sigma^{-1}\mathbf{c}(B)\}$, so the right-hand side of the inequality (4.47) decreases to zero, leaving (4.47) always satisfied. This suggests that (4.46) is less likely to hold for larger σ_B^2 .

5 Covariance-Matching Constrained Kriging

Suppose we want to predict the proportion of neighborhoods of Pittsburgh that are in PM_{10} compliance for a given day. Recall that the 24-hour standard for PM_{10} is $150 \mu\text{g}/\text{m}^3$. Consider a two-dimensional cartographic map of Pennsylvania, and denote D as some specified region in that map containing Pittsburgh. Let the i th neighborhood of D be called B_i ; $i = 1, \dots, m$, so that

$$D = \cup_{i=1}^m B_i, \text{ and } B_i \cap B_j = \emptyset; i \neq j.$$

Then the predictand describing the areally weighted proportion of neighborhoods of Pittsburgh in PM_{10} compliance can be written as,

$$F_S(t_0) = \frac{1}{|D|} \sum_{i=1}^m |B_i| I(S(B_i) \leq t_0), \quad (5.1)$$

where $S(\mathbf{u})$ denotes the true PM_{10} value at location \mathbf{u} (without measurement error), $S(B) = \frac{1}{|B|} \int_B S(\mathbf{u}) d\mathbf{u}$, $|B| = \int_B 1 ds$; $B \subset D$, and $t_0 = 150 \mu\text{g}/\text{m}^3$.

We can construct a universal constrained kriging predictor of $F_S(t_0)$ (see Section 4.2) as follows:

$$\hat{F}_{S;uc}(t_0) = \frac{1}{|D|} \sum_{i=1}^m |B_i| I(\hat{S}_{uc}(B_i) \leq t_0), \quad (5.2)$$

where $\hat{S}_{uc}(B)$ is the universal constrained kriging predictor of $S(B)$ given by (4.4). (If $S(\cdot)$ is first-order stationary, then an ordinary constrained kriging predictor of $F_S(t_0)$ may be similarly obtained.)

Denote

$$g(S(B)) \equiv I(S(B) \leq t_0).$$

Then not only is g nonlinear, it is not continuous (and hence not differentiable) when $S(B) = t_0$. Nevertheless, from Section 4.2.1, we see that if $(S(\cdot), \epsilon(\cdot))$ is a Gaussian process, then

$$g(\hat{S}_{uc}(B)) = I(\hat{S}_{uc}(B) \leq t_0),$$

is exactly unbiased for $g(S(B))$. Note that if $(S(\cdot), \epsilon(\cdot))$ is not a Gaussian process, then the unbiasedness properties of $g(S(B))$ are unclear, since we cannot appeal to the δ -method approximation for this non-smooth g . If $g(\hat{S}_{uc}(B_i))$ is unbiased for $g(S(B_i))$; $i = 1, \dots, m$, then $\hat{F}_{S;uc}(t_0)$ in (5.2), which is a weighted average of all the unbiased predictors $\{g(\hat{S}_{uc}(B_1)), \dots, g(\hat{S}_{uc}(B_m))\}$, is also unbiased.

Suppose we are interested in predicting not only the areally weighted proportion of Pittsburgh that is out of compliance, but also the *average quantity* of PM_{10} over those subregions in Pittsburgh that are out of compliance; this quantity tells us *how much* those subregions are out of compliance on average, and the predictand describing this quantity can be written as

$$Q_A(t_0) \equiv \frac{\sum_{i=1}^m |B_i| S(B_i) I(S(B_i) > t_0)}{|D|(1 - F_S(t_0))}, \quad (5.3)$$

where $F_S(t_0)$ is given by (5.1).

Observe that (5.3) *cannot* be written as

$$\sum_{i=1}^m g(S(B_i)),$$

for some g , but (5.3) can be expressed as

$$g(S) \equiv \frac{\sum_{i=1}^m |B_i| \mathbf{e}_i' S I(\mathbf{e}_i' S > t)}{\sum_{i=1}^m |B_i| I(\mathbf{e}_i' S > t)}; \quad t \in \mathbb{R}, \quad (5.4)$$

where g is a scalar-valued function that is nonlinear in its vector-valued argument,

$$S \equiv (S(B_1), \dots, S(B_m))'; \quad B_1, \dots, B_m \subset D. \quad (5.5)$$

Clearly then, equation (5.4) suggests that the constrained-kriging methodology needs to be extended to the “covariance-matching” case where multivariate predictions \hat{S}_{um} of S are required and the extra constraint becomes $\text{var}(\hat{S}_{um}) = \text{var}(S)$.

In addition, it may also be of interest to compare $Q_A(t_0)$ with the *average quantity* of PM_{10} over the remaining subregions (i.e., those that are in compliance). We denote the predictand describing this latter quantity as

$$Q_A^c(t_0) \equiv \frac{\sum_{i=1}^m |B_i| S(B_i) I(S(B_i) \leq t_0)}{|D| F_S(t_0)}. \quad (5.6)$$

5.1 Covariance-Matching Constrained Kriging Equations

Suppose we wish to predict $g(\mathbf{S})$, where g is some measurable, and for the moment, scalar-valued function. Analogous to the “univariate” constrained kriging approach described in Section 4.2, consider predictors of the form $g(\hat{\mathbf{S}})$, where $\hat{\mathbf{S}} \equiv \mathbf{A}'\mathbf{Z}$ is a linear predictor of \mathbf{S} , and \mathbf{A} is an $n \times m$ matrix.

5.1.1 Unbiasedness of $g(\mathbf{A}'\mathbf{Z})$

Assume that the coefficients $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)$ satisfy the constraints,

$$(\text{MC1}) : E(\mathbf{A}'\mathbf{Z}) = E(\mathbf{S}),$$

$$(\text{MC2}) : \text{var}(\mathbf{A}'\mathbf{Z}) = \text{var}(\mathbf{S}).$$

For future reference, we write $\mu_m \equiv E(\mathbf{S})$ and $\Sigma_m \equiv \text{var}(\mathbf{S})$, where recall that \mathbf{S} is an $m \times 1$ vector.

If $(S(\cdot), \epsilon(\cdot))$ is bivariate Gaussian, and if (MC1) and (MC2) hold, then

$$\mathbf{A}'\mathbf{Z} \stackrel{d}{=} \mathbf{S},$$

where “ $\stackrel{d}{=}$ ” denotes equality in distribution. As a consequence,

$$E(g(\mathbf{A}'\mathbf{Z})) = E(g(\mathbf{S})),$$

for any integrable function g .

Suppose that $(S(\cdot), \epsilon(\cdot))$ is not bivariate Gaussian, and suppose that g is smooth enough to possess two derivatives, that is, the $m \times 1$ vector,

$$g'(\mathbf{x}) \equiv \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}},$$

and the $m \times m$ matrix,

$$g''(\mathbf{x}) \equiv \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'},$$

both exist. Then by the δ -method (e.g. Schott, 1997; Ch. 8) we have,

$$\begin{aligned} E(g(\mathbf{A}'\mathbf{Z})) &\simeq g(E(\mathbf{A}'\mathbf{Z})) + E\{(\mathbf{A}'\mathbf{Z} - E(\mathbf{A}'\mathbf{Z}))g''(E(\mathbf{A}'\mathbf{Z}))(\mathbf{A}'\mathbf{Z} - E(\mathbf{A}'\mathbf{Z}))'\}/2 \\ &= g(E(\mathbf{A}'\mathbf{Z})) + \text{tr}\{[g''(E(\mathbf{A}'\mathbf{Z}))]\text{var}(\mathbf{A}'\mathbf{Z})\}/2, \end{aligned}$$

and similarly,

$$E(g(\mathbf{S})) \simeq g(E(\mathbf{S})) + \text{tr}\{g''(E(\mathbf{S}))\text{var}(\mathbf{S})\}/2,$$

where $\text{tr}(\cdot)$ denotes the trace operator, and $g''(\mathbf{w}) = \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \Big|_{\mathbf{x}=\mathbf{w}}$. Now, if (MC1) and (MC2) hold, then clearly,

$$E(g(\mathbf{A}'\mathbf{Z})) \simeq E(g(\mathbf{S})),$$

for g smooth enough to possess two derivatives.

5.1.2 Optimization of $g(\mathbf{A}'\mathbf{Z})$

We need to choose the $n \times m$ matrix \mathbf{A} , satisfying constraints (MC1) and (MC2), that minimizes the scalar quantity,

$$\text{MSPE}[g(\mathbf{A}'\mathbf{Z})].$$

If \mathbf{A} satisfies constraints (MC1) and (MC2) and g is smooth enough to possess two derivatives, then,

$$\begin{aligned} \text{MSPE}[g(\mathbf{A}'\mathbf{Z})] &\equiv E(g(\mathbf{A}'\mathbf{Z}) - g(\mathbf{S}))^2 \\ &\simeq \text{var}(g(\mathbf{A}'\mathbf{Z}) - g(\mathbf{S})), \end{aligned}$$

with equality holding if $(S(\cdot), \epsilon(\cdot))$ is a bivariate Gaussian process. By the δ -method (e.g. Schott, 1997; Ch. 8) we have,

$$\text{var}(g(\mathbf{A}'\mathbf{Z})) \simeq [g'(\boldsymbol{\mu}_m)]' \text{var}(\mathbf{A}'\mathbf{Z}) [g'(\boldsymbol{\mu}_m)],$$

$$\text{var}(g(\mathbf{S})) \simeq [g'(\boldsymbol{\mu}_m)]' \text{var}(\mathbf{S}) [g'(\boldsymbol{\mu}_m)],$$

and

$$\text{cov}(g(\mathbf{A}'\mathbf{Z}), g(\mathbf{S})) \simeq [g'(\boldsymbol{\mu}_m)]' \text{cov}(\mathbf{A}'\mathbf{Z}, \mathbf{S}) [g'(\boldsymbol{\mu}_m)],$$

where $g'(\boldsymbol{\mu}_m) = \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\boldsymbol{\mu}_m}$ and recall that $\boldsymbol{\mu}_m \equiv E(\mathbf{S}) = E(\mathbf{A}'\mathbf{Z})$. Therefore,

$$\text{var}(g(\mathbf{A}'\mathbf{Z}) - g(\mathbf{S})) \simeq \mathbf{y}' \mathbf{M}_A \mathbf{y},$$

where $\mathbf{y} \equiv g'(\boldsymbol{\mu}_m)$ and $\mathbf{M}_A \equiv \text{MSPE}[\mathbf{A}'\mathbf{Z}] = \text{var}(\mathbf{A}'\mathbf{Z} - \mathbf{S})$.

Thus, the objective is to choose a matrix A_0 satisfying (MC1) and (MC2) that minimizes the scalar function $y'M_A y$, or that “minimizes” the $m \times m$ matrix $M_A \equiv \text{MSPE}[A'Z]$, with respect to A . Then, from the previous paragraph, A_0 approximately minimizes the scalar quantity $\text{MSPE}[g(A'Z)]$, for all A satisfying (MC1) and (MC2).

But what is meant by “minimizing” a square symmetric matrix? There are several criteria by which such matrices are said to be minimized. The strongest criterion is that of *non-negative definiteness*, whereby a square symmetric matrix A is said to be no larger than a square symmetric matrix B of the same order, if $B - A$ is non-negative definite (n.n.d.) (i.e., all the eigenvalues of the matrix difference $(B - A)$ are non-negative). There are also weaker minimization criteria such as those that depend on the trace operator. Under these weaker criteria, a square symmetric matrix A is said to be no larger than a square symmetric matrix B of the same order, if $\text{tr}(B) \geq \text{tr}(A)$ (i.e., the sum of the eigenvalues of $(B - A)$ are non-negative).

Analogous to the univariate case in Section 4.2.3, assume that (3.10) holds and hence from (3.11), $\mu_m \equiv (\mu(B_1), \dots, \mu(B_m))' = X'_m \beta$, where $X_m \equiv (x(B_1), \dots, x(B_m))$ is a $(p+1) \times m$ matrix. Then, conditions (MC1) and (MC2) can be written as

$$(\text{MC1}) : A'X = X'_m,$$

$$(\text{MC2}) : A'\Sigma A = \Sigma_m,$$

since $A'X\beta = E(A'Z) \equiv E(S) = \mu_m = X'_m \beta$, for all β , and $A'\Sigma A = \text{var}(A'Z) \equiv \text{var}(S) = \Sigma_m$. These constraints can also be written as

$$(\text{MC1}) : a'_i X = x(B_i)'; \quad i = 1, \dots, m, \tag{5.7}$$

$$(\text{MC2}) : a'_i \Sigma a_j = C(B_i, B_j); \quad i, j = 1, \dots, m,$$

where $A = (a_1, \dots, a_m)$. Observe that (5.7) requires that the variances *and* the covariances of the elements of $A'Z$ and S match, hence the term “covariance-matching” constrained kriging.

Now,

$$\begin{aligned} y'M_A y &= y'E[(A'Z - S)(A'Z - S)']y \\ &= y'[(A'\Sigma A - A'C - C'A + \Sigma_m)y], \end{aligned}$$

where $C \equiv \text{cov}(Z, S) = (c(B_1), \dots, c(B_m))$. If A satisfies (MC2), then

$$y'M_{AY} = 2y'[\Sigma_m - A'C]y, \quad (5.8)$$

which we want to minimize with respect to A , subject to (MC1) and (MC2). Therefore, the objective function to maximize, with respect to A , is

$$\begin{aligned} f(A) &\equiv 2y'A'Cy + 2 \sum_{i=1}^m (a_i'X - x(B_i)')\lambda_{1i} - \sum_{i=1}^m \sum_{j=1}^m \lambda_{2ij}(a_i'\Sigma a_j - C(B_i, B_j)) \\ &= 2y'A'Cy + 2tr((A'X - X'_m)\Lambda_1) - tr((A'\Sigma A - \Sigma_m)\Lambda_2), \end{aligned}$$

where $\Lambda_1 \equiv (\lambda_{11}, \dots, \lambda_{1m}) = ((\lambda_{1ij}))$ (i.e., a $(p+1) \times m$ matrix whose (i, j) th element is λ_{1ij}), and $\Lambda_2 \equiv ((\lambda_{2ij}))$ (i.e., an $m \times m$ matrix whose (i, j) th element is λ_{2ij}), are Lagrange multipliers.

Define the vector of universal kriging predictors,

$$\hat{S}_{uk} \equiv (\hat{S}_{uk}(B_1), \dots, \hat{S}_{uk}(B_m))'.$$

Then,

$$\hat{S}_{uk} \equiv X'_m \hat{\beta} + C'\Sigma^{-1}(Z - X\hat{\beta}), \quad (5.9)$$

where $C = (c(B_1), \dots, c(B_m))$. Notice that \hat{S}_{uk} can be written as $A'_{uk}Z$, a linear predictor in Z , and although it satisfies (MC1) it does not satisfy (MC2).

Lemma 5.1 *Let X be an $n \times m$ matrix, A an $n \times m$ matrix, V an $n \times n$ matrix, and B an $m \times m$ matrix. If $n \neq m$, or if $n = m$ and A is not symmetric, then,*

$$\frac{\partial \text{tr}(X'A)}{\partial A} = X,$$

and

$$\frac{\partial \text{tr}(A'VAB)}{\partial A} = V'AB' + VAB.$$

Proof: See Rao (1985). □

Proposition 5.1 Assume that $P_u \equiv \text{var}(S) - \text{var}(X'_m \hat{\beta})$, $Q_u \equiv \text{var}(\hat{S}_{uk}) - \text{var}(X'_m \hat{\beta})$, and Σ are p.d., where $\hat{\beta}$ is the BLUE of β given by (3.12), and \hat{S}_{uk} is given by (5.9). Suppose that P_u and Q_u can be decomposed such that $P_{u1}^{-1} Q_{u1} y y'$ is symmetric, where $P_u = P'_{u1} P_{u1}$, $Q_u = Q'_{u1} Q_{u1}$, P_{u1} and Q_{u1} are nonsingular, and $y = g'(\mu_m)$. Then, a stationary point of the scalar function,

$$f(A) \equiv 2y'A'Cy + 2\text{tr}((A'X - X'_m)\Lambda_1) - \text{tr}(A_2(A'\Sigma A - \Sigma_m)), \quad (5.10)$$

is achieved at

$$A_0 = \Sigma^{-1}(I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1})CK_u + \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X_m \quad (5.11)$$

where $K_u \equiv Q_{u1}^{-1}P_{u1}$.

Proof: Differentiate $f(A)$ with respect to A and equate with zero to obtain

$$f'(A) = 2Cyy' + 2X\Lambda_1 - 2\Sigma A\Lambda_2^* = 0,$$

where $\Lambda_2^* = (\Lambda_2 + \Lambda_2')/2$.

Thus,

$$A\Lambda_2^* = \Sigma^{-1}(Cyy' + X\Lambda_1),$$

since Σ is p.d., and hence,

$$\Lambda_1 = (X'\Sigma^{-1}X)^{-1}[X_m\Lambda_2^* - X'\Sigma^{-1}Cyy'],$$

since (MC1) implies that $X'A = X_m$. Therefore,

$$A\Lambda_2^* = \Sigma^{-1}(I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1})Cyy' + \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X_m\Lambda_2^*. \quad (5.12)$$

Also, (MC2) implies that

$$\begin{aligned} \Lambda_2^* \Sigma_m \Lambda_2^* &\equiv \Lambda_2^* A' \Sigma A \Lambda_2^* \\ &= yy'C'\Sigma^{-1}\{\Sigma - 2X(X'\Sigma^{-1}X)^{-1}X' \\ &\quad + X(X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}X)(X'\Sigma^{-1}X)^{-1}X'\}\Sigma^{-1}Cyy' \\ &\quad + \Lambda_2^* X'_m (X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}X)(X'\Sigma^{-1}X)^{-1}X_m \Lambda_2^*, \end{aligned}$$

since $\mathbf{y}\mathbf{y}'\mathbf{C}'\{\mathbf{I} - \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\}\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m\Lambda_2^* = 0$. Therefore,

$$\Lambda_2^*\Sigma_m\Lambda_2^* = \mathbf{y}\mathbf{y}'(\mathbf{C}'\Sigma^{-1}\mathbf{C} - \mathbf{C}'\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{C})\mathbf{y}\mathbf{y}' + \Lambda_2^*\mathbf{X}_m'(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m\Lambda_2^*.$$

That is,

$$\Lambda_2^*\mathbf{P}_u\Lambda_2^* = \mathbf{y}\mathbf{y}'\mathbf{Q}_u\mathbf{y}\mathbf{y}', \quad (5.13)$$

where

$$\mathbf{P}_u \equiv \text{var}(\mathbf{S}) - \text{var}(\mathbf{X}_m'\hat{\boldsymbol{\beta}}) = \Sigma_m - \mathbf{X}_m'(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m,$$

and

$$\mathbf{Q}_u \equiv \text{var}(\hat{\mathbf{S}}_{uk}) - \text{var}(\mathbf{X}_m'\hat{\boldsymbol{\beta}}) = \mathbf{C}'\Sigma^{-1}\mathbf{C} - \mathbf{C}'\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{C}.$$

If \mathbf{P}_u and \mathbf{Q}_u are p.d., then there exist (non-unique) nonsingular matrices \mathbf{P}_{u1} and \mathbf{Q}_{u1} such that $\mathbf{P}_u = \mathbf{P}_{u1}'\mathbf{P}_{u1}$ and $\mathbf{Q}_u = \mathbf{Q}_{u1}'\mathbf{Q}_{u1}$. Now, by assumption, \mathbf{P}_u and \mathbf{Q}_u can be decomposed such that $\mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1}\mathbf{y}\mathbf{y}'$ is symmetric. Then

$$\Lambda_2^* = \mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1}\mathbf{y}\mathbf{y}', \quad (5.14)$$

is a solution to (5.13). Also, from (5.12) and (5.14),

$$\mathbf{A}\Lambda_2^* = \mathbf{L}\mathbf{y}\mathbf{y}',$$

where

$$\mathbf{L} = \Sigma^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1})\mathbf{C} + \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m\mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1}.$$

Therefore, from (5.14), $(\mathbf{A}\mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1} - \mathbf{L})\mathbf{y}\mathbf{y}' = 0$, for all \mathbf{y} , and this implies that $\text{tr}\{(\mathbf{A}\mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1} - \mathbf{L})\mathbf{y}\mathbf{y}'\} = 0$, for all \mathbf{y} . Hence, $\mathbf{y}'(\mathbf{A}\mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1} - \mathbf{L})\mathbf{y} = 0$, for all \mathbf{y} , which implies that $\mathbf{A}\mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1} - \mathbf{L} = 0$. Therefore,

$$\mathbf{A}_0 = \Sigma^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1})\mathbf{C}\mathbf{Q}_{u1}^{-1}\mathbf{P}_{u1} + \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m,$$

where $f(\mathbf{A}_0)$ is a stationary point. □

Proposition 5.1 shows that under certain conditions, we may obtain a matrix A_0 , given by (5.11), for which the objective function $f(A)$, given by (5.10), has zero derivative at A_0 . Suppose we construct a predictor $A'_0 Z$ of S . Will this predictor be optimal in any sense? Lemma 5.2, Proposition 5.2, and Proposition 5.3 below, delineate the precise conditions under which $A'_0 Z$ is optimal.

Lemma 5.2 *Let A be any $n \times m$ matrix satisfying (MC1) and (MC2), and consider A_0 given by (5.11). Then,*

$$(A_0 - A)'C = (\Sigma_m - A'\Sigma A_0)K_u^{-1}, \quad (5.15)$$

where $K_u^{-1} \equiv P_{u1}^{-1}Q_{u1}$, P_u and Q_u are arbitrarily decomposed as $P_u = P'_{u1}P_{u1}$ and $Q_u = Q'_{u1}Q_{u1}$, and P_{u1} and Q_{u1} are nonsingular.

Proof: Since (MC1) implies that $A'X = X_m$, we have

$$\begin{aligned} A'\Sigma A_0 K_u^{-1} &= A'(I - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1})C + A'X(X'\Sigma^{-1}X)^{-1}X_m K_u^{-1} \\ &= A'C + X'_m(X'\Sigma^{-1}X)^{-1}(X_m K_u^{-1} - X'\Sigma^{-1}C), \end{aligned}$$

and hence,

$$A'C = A'\Sigma A_0 K_u^{-1} - X'_m(X'\Sigma^{-1}X)^{-1}(X_m K_u^{-1} - X'\Sigma^{-1}C).$$

Now,

$$A'_0 C = K'_u C'(I - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X')\Sigma^{-1}C + X'_m(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}C.$$

Therefore,

$$\begin{aligned} (A_0 - A)'C &= K'_u C'(I - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X')\Sigma^{-1}C \\ &\quad - (A'\Sigma A_0 - X'_m(X'\Sigma^{-1}X)^{-1}X_m)K_u^{-1} \\ &= P'_{u1}(Q_{u1}^{-1})'Q_u - (A'\Sigma A_0 - \Sigma_m + P_u)K_u^{-1}, \end{aligned}$$

since $P_u \equiv \Sigma_m - X'_m(X'\Sigma^{-1}X)^{-1}X_m$ and $Q_u \equiv C'\Sigma^{-1}C - C'X(X'\Sigma^{-1}X)^{-1}XC$. Therefore,

$$\begin{aligned} (A_0 - A)'C &= P'_{u1}Q_{u1} + (\Sigma_m - A'\Sigma A_0)K_u^{-1} - P_u P_{u1}^{-1}Q_{u1} \\ &= (\Sigma_m - A'\Sigma A_0)K_u^{-1}, \end{aligned}$$

giving the desired result. \square

Proposition 5.2 *Let A be any $n \times m$ matrix satisfying (MC1) and (MC2), and consider A_0 given by (5.11). Define $M_A \equiv E(S - A'Z)(S - A'Z)'$ and likewise for M_{A_0} . Then,*

$$M_A - M_{A_0} = (\Sigma_m - A'\Sigma A_0)K_u^{-1} + (K_u^{-1})'(\Sigma_m - A_0'\Sigma A), \quad (5.16)$$

where $K_u^{-1} \equiv P_{u1}^{-1}Q_{u1}$, P_u and Q_u are arbitrarily decomposed as $P_u = P_{u1}'P_{u1}$ and $Q_u = Q_{u1}'Q_{u1}$, and P_{u1} and Q_{u1} are nonsingular.

Proof: From Proposition 5.4, A_0 satisfies (MC2), and hence $A_0'\Sigma A_0 = \Sigma_m$. Therefore, Lemma 5.2 implies that

$$\begin{aligned} M_A - M_{A_0} &= (2\Sigma_m - A'C - C'A) - (2\Sigma_m - A_0'C - C'A_0) \\ &= (A_0 - A)'C + C'(A_0 - A) \\ &= (\Sigma_m - A'\Sigma A_0)K_u^{-1} + (K_u^{-1})'(\Sigma_m - A_0'\Sigma A), \end{aligned}$$

giving the desired result. \square

Proposition 5.3 *Let A be any $n \times m$ matrix satisfying (MC1) and (MC2), and consider A_0 given by (5.11). Suppose that $P_u = P_{u1}'P_{u1}$, $Q_u = Q_{u1}'Q_{u1}$, and $K_u^{-1} = P_{u1}^{-1}Q_{u1}$ are such that $K_u^{-1}yy' \equiv P_{u1}^{-1}Q_{u1}yy'$ is symmetric. Further suppose that*

$$K_u^{-1} + (K_u^{-1})' \text{ is p.d.} \quad (5.17)$$

Then $y'M_A y \geq y'M_{A_0} y$ for every A satisfying (MC1) and (MC2), where $y \equiv g'(\mu_m) \neq 0$.

Proof: By supposition, $K_u^{-1} + (K_u^{-1})'$ is p.d. Then,

$$\begin{aligned} \text{tr}(K_u^{-1}yy') &= \text{tr}(K_u^{-1}yy' + (K_u^{-1}yy')')/2 \\ &= y'(K_u^{-1} + (K_u^{-1})')y/2 \\ &> 0. \end{aligned}$$

Now, $\text{rank}(\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}') \leq \text{rank}(\mathbf{y}\mathbf{y}') = 1$. Thus, $\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}'$ has at most one nonzero eigenvalue, and hence that eigenvalue is equal to $\text{tr}(\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}') > 0$. Therefore, $\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}'$ is n.n.d., and hence can be written as $\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}' = \mathbf{H}'\mathbf{H}$, for some matrix \mathbf{H} .

Since \mathbf{A} and \mathbf{A}_0 satisfy (MC2), we have

$$\mathbf{M}_A - \mathbf{M}_{A_0} = (\mathbf{A}_0 - \mathbf{A})'\mathbf{C} + \mathbf{C}'(\mathbf{A}_0 - \mathbf{A}),$$

and from Lemma 5.2, we have

$$\begin{aligned} \mathbf{y}'(\mathbf{M}_A - \mathbf{M}_{A_0})\mathbf{y} &= \text{tr}[(\mathbf{A}_0 - \mathbf{A})'\mathbf{C}\mathbf{y}\mathbf{y}' + \mathbf{y}\mathbf{y}'\mathbf{C}'(\mathbf{A}_0 - \mathbf{A})] \\ &= \text{tr}[(\Sigma_m - \mathbf{A}'\Sigma\mathbf{A}_0)\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}' + \mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}'(\Sigma_m - \mathbf{A}_0'\Sigma\mathbf{A})] \\ &= \text{tr}[\mathbf{H}'(\Sigma_m - \mathbf{A}'\Sigma\mathbf{A}_0)\mathbf{H} + \mathbf{H}'(\Sigma_m - \mathbf{A}_0'\Sigma\mathbf{A})\mathbf{H}]. \end{aligned}$$

Therefore, since $\Sigma_m = \mathbf{A}'\Sigma\mathbf{A} = \mathbf{A}_0'\Sigma\mathbf{A}_0$, and Σ is p.d., we obtain

$$\mathbf{y}'(\mathbf{M}_A - \mathbf{M}_{A_0})\mathbf{y} = \text{tr}[\mathbf{H}'(\mathbf{A}_0 - \mathbf{A})'\Sigma(\mathbf{A}_0 - \mathbf{A})\mathbf{H}] \geq 0,$$

giving the desired result. \square

That is, if all the suppositions in Proposition 5.1 hold, then $f(\mathbf{A}_0)$ has zero derivative. If, in addition, $\mathbf{K}_u^{-1} + (\mathbf{K}_u^{-1})'$ is p.d., then the predictor $\mathbf{A}_0'\mathbf{Z}$ is optimal.

Now, suppose that \mathbf{P}_u and \mathbf{Q}_u are p.d., but that they are decomposed such that $\mathbf{K}_u^{-1}\mathbf{y}\mathbf{y}' = \mathbf{P}_{u1}^{-1}\mathbf{Q}_{u1}\mathbf{y}\mathbf{y}'$ is *not* symmetric. This implies that Λ_2^* in (5.14) may no longer be a solution to (5.13). What are the consequences of this? Proposition 5.4 below shows that in such a case, even though $f(\mathbf{A}_0)$ may no longer have zero derivative for \mathbf{A}_0 given by (5.11), \mathbf{A}_0 still satisfies the constraints (MC1) and (MC2).

Proposition 5.4 *Suppose that $\mathbf{P}_u \equiv \text{var}(\mathbf{S}) - \text{var}(\mathbf{X}_m'\hat{\beta})$ and $\mathbf{Q}_u \equiv \text{var}(\hat{\mathbf{S}}_{uk}) - \text{var}(\mathbf{X}_m'\hat{\beta})$ are p.d. and are decomposed such that $\mathbf{P}_u = \mathbf{P}_{u1}'\mathbf{P}_{u1}$ and $\mathbf{Q}_u = \mathbf{Q}_{u1}'\mathbf{Q}_{u1}$, where \mathbf{P}_{u1} and \mathbf{Q}_{u1} are nonsingular matrices. Then \mathbf{A}_0 given by (5.11), namely,*

$$\mathbf{A}_0 = \Sigma^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1})\mathbf{C}\mathbf{Q}_{u1}^{-1}\mathbf{P}_{u1} + \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m,$$

satisfies the constraints (MC1) and (MC2).

Proof: The constraint (MC1) is satisfied, since

$$\begin{aligned} A'_0 X &= K'_u C' (I - \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} X') \Sigma^{-1} X + X'_m (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} X \\ &= X'_m. \end{aligned}$$

Further,

$$A'_0 \Sigma A_0 = K'_u C' \Sigma^{-1} (I - X (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1}) C K_u + X'_m (X' \Sigma^{-1} X)^{-1} X_m,$$

since $(I - \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} X') \Sigma^{-1} X' = 0$. And, since $P_u = \Sigma_m - X'_m (X' \Sigma^{-1} X)^{-1} X_m$ and $Q_u = C' \Sigma^{-1} C - C' \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} C$, we have

$$\begin{aligned} A'_0 \Sigma A_0 &= K'_u Q_u K_u + \Sigma_m - P_u \\ &= P'_{u1} (Q_{u1}^{-1})' Q_u Q_{u1}^{-1} P_{u1} + \Sigma_m - P_u \\ &= \Sigma_m. \end{aligned}$$

That is, (MC2) is satisfied. □

Proposition 5.4 shows that for *any* nonsingular decomposition-matrices P_{u1} and Q_{u1} , A_0 still satisfies (MC1) and (MC2). Now, suppose that we cannot find P_{u1} and Q_{u1} such that $P_{u1}^{-1} Q_{u1} y y'$ is symmetric. How do we decompose P_u and Q_u so that the matrix $MSPE(A'_0 Z)$ is as “small” as possible? Proposition 5.5 in Section 5.1.3 suggests how to decompose P_u and Q_u for that purpose.

5.1.3 Covariance-Matching Constrained Kriging Predictor

Suppose that P_u and Q_u are p.d. and that they can be decomposed such that the suppositions given in Proposition 5.1 and Proposition 5.3 are satisfied. Define the *universal covariance-matching constrained kriging predictor* of S as

$$\hat{S}_{um} \equiv A'_0 Z = X'_m \hat{\beta} + K'_u C' \Sigma^{-1} (Z - X \hat{\beta}), \quad (5.18)$$

where $\hat{\beta} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Z$ is the BLUE of β given by (3.12), and $K_u = Q_{u1}^{-1} P_{u1}$. Then, Proposition 5.1 and Proposition 5.3 imply that \hat{S}_{um} is an optimal predictor of S with respect to squared-error loss.

Proposition 5.5 *The MSPE of \hat{S}_{um} in (5.18) is given by*

$$\begin{aligned} \text{MSPE}(\hat{S}_{um}) &= 2\Sigma_m - \mathbf{X}'_m(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{C} - \mathbf{C}'\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_m \\ &\quad - (\mathbf{P}'_{u1}\mathbf{Q}_{u1} + \mathbf{Q}'_{u1}\mathbf{P}_{u1}). \end{aligned} \quad (5.19)$$

Proof: The MSPE of \hat{S}_{um} is given by

$$\begin{aligned} \text{MSPE}(\hat{S}_{um}) &\equiv E[(\hat{S}_{um} - \mathbf{S})(\hat{S}_{um} - \mathbf{S})'] \\ &= 2\Sigma_m - \text{cov}(\hat{S}_{um}, \mathbf{S}) - \text{cov}(\mathbf{S}, \hat{S}_{um}), \end{aligned}$$

where

$$\begin{aligned} \text{cov}(\hat{S}_{um}, \mathbf{S}) &= \mathbf{X}'_m(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{C} \\ &\quad + \mathbf{K}'_u\{\mathbf{C}'\Sigma^{-1}\mathbf{C} - \mathbf{C}'\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{C}\} \\ &= \mathbf{X}'_m(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{C} + \mathbf{P}'_{u1}\mathbf{Q}_{u1}, \end{aligned}$$

and the result follows. \square

Suppose that decompositions of \mathbf{P}_u and \mathbf{Q}_u cannot be found that satisfy the suppositions of Proposition 5.1 and Proposition 5.3. Then, (5.19) suggests that \mathbf{P}_u and \mathbf{Q}_u be decomposed so that $(\mathbf{P}'_{u1}\mathbf{Q}_{u1} + \mathbf{Q}'_{u1}\mathbf{P}_{u1})$ is as “large” as possible.

Now, consider the special case where $S(\cdot)$ is first-order stationary (i.e., where $\mathbf{x}(s) \equiv 1$). Define the vector of ordinary kriging predictors,

$$\hat{S}_{ok} \equiv (\hat{S}_{ok}(B_1), \dots, \hat{S}_{ok}(B_m))'.$$

Then,

$$\hat{S}_{ok} \equiv \hat{\mu}\mathbf{1}_m + \mathbf{C}'\Sigma^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1}), \quad (5.20)$$

where $\hat{\mu}$ is the BLUE of μ given by (3.15) and $\mathbf{1}_m$ is a vector of m ones. Assume that $\mathbf{P}_o = \mathbf{P}'_{o1}\mathbf{P}_{o1} \equiv \text{var}(\mathbf{S}) - \text{var}(\hat{\mu}\mathbf{1}_m)$ and $\mathbf{Q}_o = \mathbf{Q}'_{o1}\mathbf{Q}_{o1} \equiv \text{var}(\hat{S}_{ok}) - \text{var}(\hat{\mu}\mathbf{1}_m)$ are p.d., with the decompositions chosen such that $\mathbf{P}_o^{-1}\mathbf{Q}_{o1}\mathbf{y}\mathbf{y}'$ is symmetric, and $\mathbf{K}_o^{-1} + (\mathbf{K}_o^{-1})'$ is p.d. Then the *ordinary covariance-matching constrained kriging predictor* of \mathbf{S} is defined as

$$\hat{S}_{om} \equiv \hat{\mu}\mathbf{1}_m + \mathbf{K}'_o\mathbf{C}'\Sigma^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1}), \quad (5.21)$$

where

$$\mathbf{K}_o \equiv \mathbf{Q}_{o1}^{-1} \mathbf{P}_{o1}. \quad (5.22)$$

The MSPE of $\hat{\mathbf{S}}_{om}$ is given by

$$\text{MSPE}(\hat{\mathbf{S}}_{om}) = 2\mathbf{\Sigma}_m - (\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}(\mathbf{1}_m\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{C} + \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{1}\mathbf{1}_m') - (\mathbf{P}_{o1}'\mathbf{Q}_{o1} + \mathbf{Q}_{o1}'\mathbf{P}_{o1}). \quad (5.23)$$

5.1.4 Existence of the Covariance-Matching Constrained Kriging Predictor

Examination of Proposition 5.1 shows that \mathbf{A}_0 exists only if

$$(\text{ME1}) : \mathbf{Q}_u \equiv \text{var}(\hat{\mathbf{S}}_{uk}) - \text{var}(\mathbf{X}_m'\hat{\boldsymbol{\beta}}) \text{ is p.d.,}$$

$$(\text{ME2}) : \mathbf{P}_u \equiv \text{var}(\mathbf{S}) - \text{var}(\mathbf{X}_m'\hat{\boldsymbol{\beta}}) \text{ is p.d.}$$

Proposition 5.6 *The matrix \mathbf{Q}_u is n.n.d.*

Proof: The matrix,

$$\begin{aligned} \mathbf{Q}_u &\equiv \text{var}(\hat{\mathbf{S}}_{uk}) - \text{var}(\mathbf{X}_m'\hat{\boldsymbol{\beta}}) \\ &= \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C} - \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{C} \\ &= \mathbf{H}'\mathbf{H}, \end{aligned}$$

where $\mathbf{H} = \mathbf{P}\mathbf{R}\mathbf{C}$. since $\mathbf{\Sigma}$ p.d. implies that $\mathbf{\Sigma}^{-1} \equiv \mathbf{R}'\mathbf{R}$ for some matrix \mathbf{R} , and $\mathbf{P} \equiv (\mathbf{I} - \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}')$ is symmetric and idempotent. \square

Thus, existence condition (ME1) can only be violated if \mathbf{Q}_u possesses any zero-valued eigenvalues. Should this occur, a practical solution to this problem might be to add a small positive constant to the zero-valued eigenvalues of \mathbf{Q}_u . Unfortunately, such a solution may not be suitable if existence condition (ME2) is violated, since

$$\mathbf{P}_u \equiv \text{var}(\mathbf{S}) - \text{var}(\mathbf{X}_m'\hat{\boldsymbol{\beta}}) = \mathbf{\Sigma}_m - \mathbf{X}_m'(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m$$

may not be n.n.d. (i.e., it may possess negative eigenvalues).

If (ME2) is violated, a solution may possibly be found by partitioning the prediction vector S into subvectors and relaxing the covariance-matching constraints so that they apply only to elements *within*, but not *between*, partitions. For example, suppose that S is partitioned into r parts,

$$S \equiv [S'_1 | \dots | S'_r]'$$

Then, the “partitioned” constraints may be written as

$$(\text{PMC1}) : E(A'_i Z) = E(S_i); \quad i = 1, \dots, r,$$

$$(\text{PMC2}) : \text{var}(A'_i Z) = \text{var}(S_i); \quad i = 1, \dots, r,$$

where A is partitioned conformably as

$$A \equiv [A_1 | \dots | A_r].$$

Note that (PMC1) is exactly equivalent to (MC1), but that (PMC2) is different from (MC2).

Proposition 5.1 can be applied to the prediction of each of the subvectors, S_1, \dots, S_r , and if in each case the predictors exist, then we obtain the “*partitioned*” *universal covariance-matching constrained kriging predictor* of S , given by

$$\hat{S}_{up} \equiv X'_m \hat{\beta} + K'_{up} C' \Sigma^{-1} (Z - X \hat{\beta}), \quad (5.24)$$

where $K_{up} \equiv Q_{up1}^{-1} P_{up1}$. In (5.24), $P_{up} \equiv P'_{up1} P_{up1}$ is a block-diagonal matrix whose i th block comprises the submatrix

$$P_{up,i} \equiv P'_{up1,i} P_{up1,i} = \text{var}(S_i) - \text{var}(X'_{m,i} \hat{\beta}); \quad i = 1, \dots, r,$$

where $X_m \equiv [X_{m,1} | \dots | X_{m,r}]$. Similarly, $Q_{up} \equiv Q'_{up1} Q_{up1}$ is a block-diagonal matrix whose i th block comprises

$$Q_{up,i} \equiv Q'_{up1,i} Q_{up1,i} = \text{var}(\hat{S}_{uk,i}) - \text{var}(X'_{m,i} \hat{\beta}); \quad i = 1, \dots, r,$$

where $\hat{S}_{uk} \equiv [\hat{S}'_{uk,1} | \dots | \hat{S}'_{uk,r}]'$. That is, the elements in the diagonal blocks of P_{up} are identical to the corresponding elements of P_u , but all other elements of P_{up} are zero; similarly for Q_{up} .

If necessary, the partitioning of \mathbf{S} could be refined to the point where each partition contains only one element, that is,

$$\mathbf{S} \equiv [S(B_1) | \dots | S(B_m)]'.$$

Then, \mathbf{P}_{up} is a diagonal matrix with i th diagonal entry,

$$\mathbf{P}_{up,i} \equiv \mathbf{P}'_{up1,i} \mathbf{P}_{up1,i} = \text{var}(S(B_i)) - \text{var}(\mathbf{x}(B_i)' \hat{\boldsymbol{\beta}}); \quad i = 1, \dots, m,$$

and \mathbf{Q}_{up} is a diagonal matrix with i th diagonal entry,

$$\mathbf{Q}_{up,i} \equiv \mathbf{Q}'_{up1,i} \mathbf{Q}_{up1,i} = \text{var}(\hat{S}_{uk}(B_i)) - \text{var}(\mathbf{x}(B_i)' \hat{\boldsymbol{\beta}}); \quad i = 1, \dots, m.$$

In this case,

$$\hat{\mathbf{S}}_{up} \equiv (\hat{S}_{uc}(B_1), \dots, \hat{S}_{uc}(B_m))',$$

where $\hat{S}_{uc}(B)$ is the universal constrained kriging predictor of $S(B)$ given by (4.4).

5.2 Summary

Suppose that $\mathbf{P}_u \equiv \text{var}(\mathbf{S}) - \text{var}(\mathbf{X}'_m \hat{\boldsymbol{\beta}})$ and $\mathbf{Q}_u \equiv \text{var}(\hat{\mathbf{S}}_{uk}) - \text{var}(\mathbf{X}'_m \hat{\boldsymbol{\beta}})$ can be decomposed so that the suppositions of Proposition 5.1 and Proposition 5.3 are satisfied. Then the covariance-matching constrained kriging predictor (5.18) of \mathbf{S} is optimal with respect to squared-error loss. In practice, it may not be easy to find decompositions of \mathbf{P}_u and \mathbf{Q}_u such that the suppositions of Proposition 5.1 and Proposition 5.3 are satisfied. However, Proposition 5.5 suggests that a suboptimal solution may be found by “maximizing” the matrix products $\mathbf{P}'_{u1} \mathbf{Q}_{u1} + \mathbf{Q}'_{u1} \mathbf{P}_{u1}$.

But, (5.18) exists only if \mathbf{P}_u and \mathbf{Q}_u are p.d. What if either \mathbf{P}_u or \mathbf{Q}_u is not p.d.? In Section 5.1.4, we showed that \mathbf{Q}_u is n.n.d. Therefore, if \mathbf{Q}_u is not p.d., a practical solution may be to replace the zero-valued eigenvalues with small positive values. On the other hand, \mathbf{P}_u may possess negative eigenvalues, and if so, the only solution may be to partition \mathbf{S} and its corresponding predictor into subvectors, and match covariances between, but not within, the corresponding subvectors as described in Section 5.1.4. Suppose that a partition of \mathbf{S} exists such that \mathbf{P}_{up} in (5.24) is p.d. Then, we can obtain the “partitioned” predictor (5.24) of \mathbf{S} .

Clearly, such a partition of S is unlikely to be unique. How will the choice of partition (for which P_{up} is p.d.) affect the resulting predictor?

6 Computer-Simulation Experiment

Consider a spatial domain over which a spatial model is defined. A computer-simulation experiment was devised to explore how covariance-matching constrained kriging compares with ordinary kriging and constrained kriging in predicting certain linear and nonlinear functionals of the state process $S(\cdot)$ defined over the spatial domain in question, under different conditions. The details of this experiment are now presented.

6.1 Spatial Domain and Spatial Model

In this experiment, we consider a square domain in \mathbb{R}^2 , discretized into $D = \{(x, y) : x, y = 1, \dots, 10\}$, a square 10×10 grid of 100 locations; see Figure 6.1. Two subdomains, $B_1 = \{(2, 2), (2, 3), (3, 2), (3, 3)\}$ and $B_2 = \{(7, 9), (7, 10), (8, 9), (8, 10)\}$ (i.e., each consisting of a 2×2 grid of 4 locations), are given special emphasis in the simulation, and their locations are identified as “1” and “2”, respectively, in Figure 6.1. In addition, the 12 locations $\{s_1, \dots, s_{12}\}$ in D marked with a circle in Figure 6.1 represent the fixed set A of sampling locations.

The spatial model over D is defined as

$$Z(s) \equiv S(s) + \epsilon(s); \quad s = (x, y)' \in D, \quad (6.1)$$

where $\epsilon(\cdot)$, independent of $S(\cdot)$, is a zero-mean, Gaussian, white-noise measurement-error process such that $\text{var}(\epsilon(s)) \equiv \tau^2$, and τ^2 is specified in Section 6.2 below. The state process is also specified in Section 6.2.

Given the sampling locations $A \equiv \{s_1, \dots, s_{12}\}$, the data $\mathbf{Z} \equiv (Z(s_1), \dots, Z(s_{12}))'$ were obtained by generating 12 ϵ -values according to model (6.1) and adding these values to the corresponding S -values generated according the appropriate state-process model given in Section 6.2.

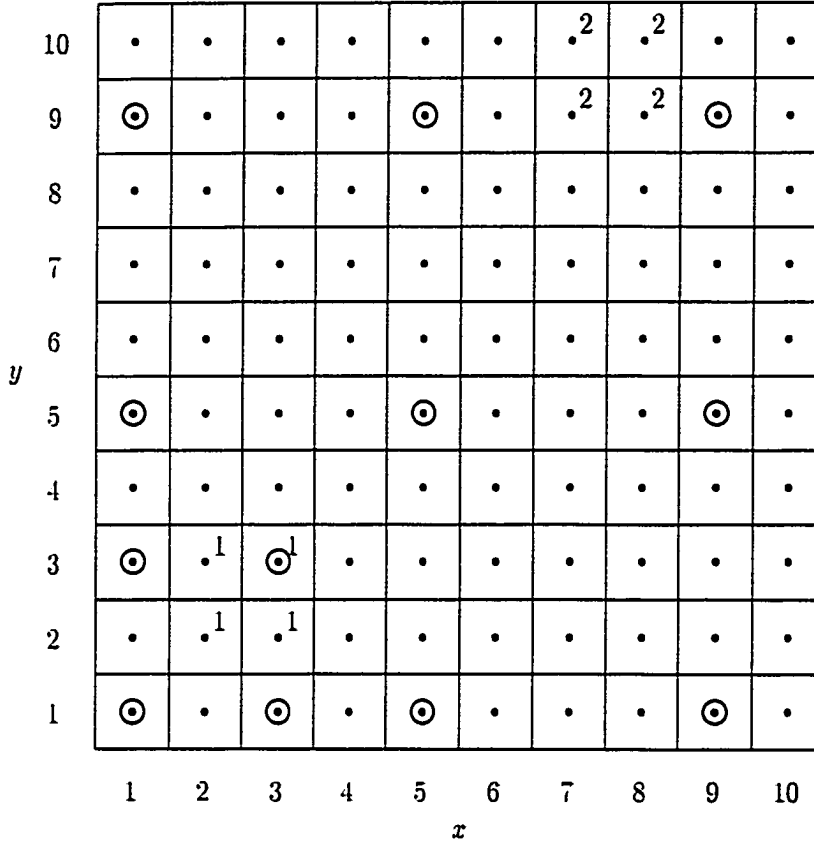


Figure 6.1 Domain D comprises a square 10×10 grid of 100 locations. The subregions, B_1 and B_2 , each comprise a 2×2 grid of 4 locations, and their locations are identified as “1” and “2”, respectively. The 12 locations marked with a circle represent the sample A .

6.2 Factors of the Experiment

Four factors were considered in this experiment: (i) distribution of the random field $S(\cdot)$, (ii) strength of spatial correlation, (iii) noise (i.e., measurement error), (iv) predictor.

Three random fields $\{S_i(s) : s \in D\}$; $i = 1, 2, 3$, were considered. In the first case,

$$S_1(s) \equiv 10 + \delta(s); \quad s = (x, y)' \in D, \quad (6.2)$$

where $\delta(\cdot)$ is a zero-mean, Gaussian spatial process with “spherical” covariance function, given by

$$C_1(\mathbf{h}) \equiv \text{cov}(\delta(s), \delta(s + \mathbf{h}))$$

$$= \sigma^2 \begin{cases} 1 - \frac{3}{2}(\|\mathbf{h}\|/a) + \frac{1}{2}(\|\mathbf{h}\|/a)^3; & \|\mathbf{h}\| \leq a \\ 0; & \|\mathbf{h}\| > a, \end{cases}$$

$\sigma^2 = 1$, and the range parameter a (specified below) governs the strength of spatial correlation.

The second random field is a log Gaussian process, defined as

$$S_2(\mathbf{s}) \equiv k_1 + k_2 \exp(S_1(\mathbf{s})); \quad \mathbf{s} = (x, y)' \in D,$$

where k_1 and k_2 are chosen so that the mean and variance of $S_2(\mathbf{s})$ are equal to those of $S_1(\mathbf{s})$; $\mathbf{s} \in D$. The third random field is defined as

$$S_3(\mathbf{s}) \equiv k_3 - k_4 \exp(S_1(\mathbf{s})); \quad \mathbf{s} = (x, y)' \in D.$$

That is, the distribution of $S_3(\mathbf{s})$ is a “mirror image” of that of $S_2(\mathbf{s})$, and the constants k_3 and k_4 are chosen so that the mean and variance of $S_3(\mathbf{s})$ are equal to those of $S_1(\mathbf{s})$ and $S_2(\mathbf{s})$; $\mathbf{s} \in D$. From Cressie (1993a, Sect. 3.2.2), it is easy to show that $k_1 = 10 - k_2 e^{10.5}$, $k_2 = (e^{21}(e - 1))^{-\frac{1}{2}}$, $k_3 = 10 + k_4 e^{10.5}$, and $k_4 = k_2$. Observe that the marginal distribution of $S_1(\mathbf{s})$ is symmetric, that of $S_2(\mathbf{s})$ is heavily skewed to the right, and that of $S_3(\mathbf{s})$ is heavily skewed to the left. This is illustrated in the top row of histograms in Figure 6.2 for which 1000 realizations from the random fields $S_1(\cdot)$, $S_2(\cdot)$, and $S_3(\cdot)$ were generated.

The strength of spatial correlation in model (6.2) is governed by the range parameter a . Three levels of a were chosen: $a = 4.1$, 7.5 , and 30 . These values represent weak, moderate, and strong correlation structures, respectively, for $S_1(\cdot)$. For $S_2(\cdot)$ and $S_3(\cdot)$, a given value of a provides a weaker correlation structure than it does for $S_1(\cdot)$. From Cressie (1993a, Sect. 3.2.2), we see that

$$\begin{aligned} C_2(\mathbf{h}) &\equiv \text{cov}(S_2(\mathbf{s}), S_2(\mathbf{s} + \mathbf{h})) \\ &= k_2^2 \text{cov}(\exp(S_1(\mathbf{s})), \exp(S_1(\mathbf{s} + \mathbf{h}))) \\ &= k_2^2 \exp\{(E(S_1(\mathbf{s})) + \text{var}(S_1(\mathbf{s}))/2)(E(S_1(\mathbf{s} + \mathbf{h})) + \text{var}(S_1(\mathbf{s} + \mathbf{h}))/2)\} \\ &\quad \times (\exp\{C_1(\mathbf{h})\} - 1) \\ &= (e^{C_1(\mathbf{h})} - 1)/(e - 1) \\ &\leq C_1(\mathbf{h}), \end{aligned}$$

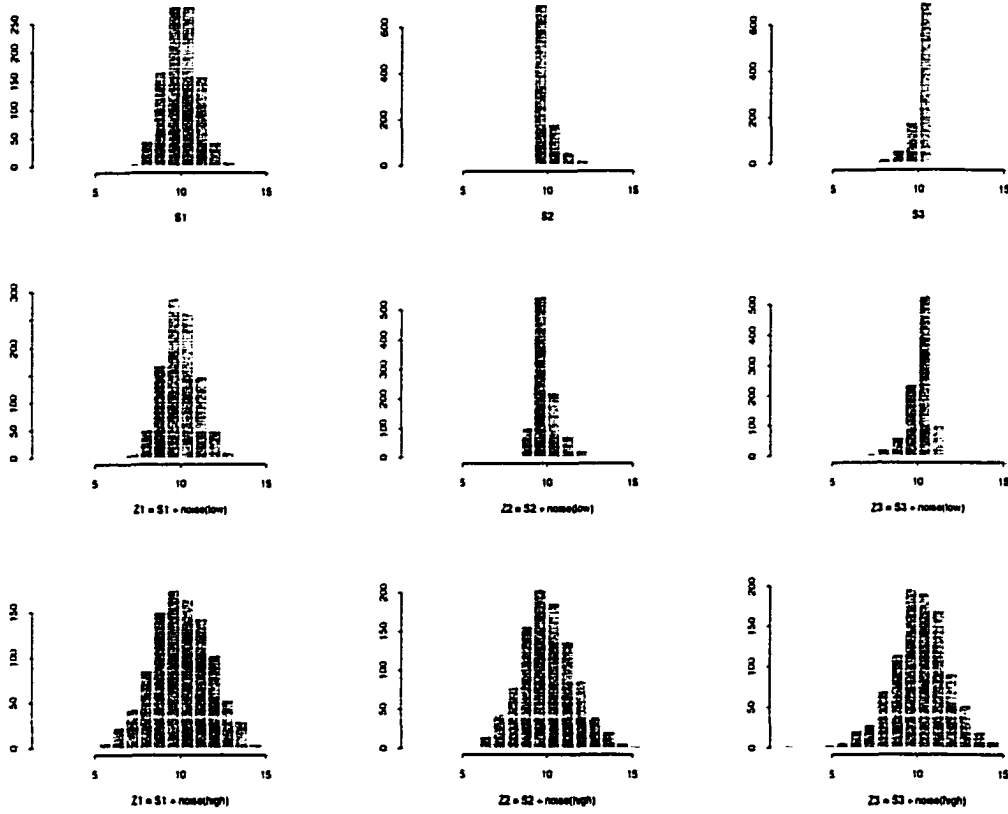


Figure 6.2 Histograms of the marginal distributions of $S_1(s)$, $S_2(s)$, and $S_3(s)$ are given in the top row of plots. In the second row, low-level noise is added to the state processes, and in the third row, high-level noise is added to the state processes. These histograms were formed by generating 1000 realizations from the appropriate distributions.

since $(e^x - 1)/(e - 1) \leq x$; $0 \leq x \leq 1$. It is easy to show that $C_3(\mathbf{h}) \equiv \text{cov}(S_3(s), S_3(s + \mathbf{h})) = C_2(\mathbf{h})$. Table 6.1 gives the parameter values chosen, along with the covariances $C_1^o(\|\mathbf{h}\|)$ and $C_2^o(\|\mathbf{h}\|)$; $\|\mathbf{h}\| = 1, \dots, 10$, where $C_1^o(\|\mathbf{h}\|) \equiv C_1(\mathbf{h})$ and $C_2^o(\|\mathbf{h}\|) \equiv C_2(\mathbf{h})$. The value $a = 4.1$ was chosen to represent the weakest level of spatial correlation, because that was the smallest value of a at which one of the predictors (presented below) could be defined. The value $a = 30$ was chosen to represent the strongest level of spatial correlation; see Table 6.1, where the strength of spatial correlation is given as a function of lag distance. An intermediate level of spatial correlation was represented by the value $a = 7.5$.

Table 6.1 Three values of the range parameter a , and the covariances $C_1^o(\|\mathbf{h}\|)$ and $C_2^o(\|\mathbf{h}\|)$; $\|\mathbf{h}\| = 1, \dots, 10$, are presented.

	$C_1^o(\ \mathbf{h}\)$			$C_2^o(\ \mathbf{h}\)$		
	a			a		
$\ \mathbf{h}\ $	4.1	7.5	30	4.1	7.5	30
1	0.641	0.801	0.950	0.523	0.715	0.923
2	0.326	0.609	0.900	0.225	0.489	0.850
3	0.098	0.432	0.850	0.060	0.314	0.780
4	0.001	0.276	0.801	0.001	0.185	0.715
5	0.000	0.148	0.752	0.000	0.093	0.653
6	0.000	0.056	0.704	0.000	0.034	0.595
7	0.000	0.007	0.656	0.000	0.004	0.540
8	0.000	0.000	0.609	0.000	0.000	0.489
9	0.000	0.000	0.564	0.000	0.000	0.440
10	0.000	0.000	0.519	0.000	0.000	0.395

Two levels of measurement error were considered: $\tau^2 = 0.1$ and 2. These two levels will be referred to as *low noise* and *high noise*, respectively. The two values of τ^2 were chosen according to the square-root signal-to-noise ratio scale, σ/τ , from which we obtain the two ratios: $\sigma/\tau \in \{0.71, 3.16\}$. In the second and third rows of plots in Figure 6.2, histograms of the marginal distributions of the three S -processes with added noise at both low and high levels, are displayed. Note that the noise-contaminated marginal distributions of $Z_2(\mathbf{s})$ and $Z_3(\mathbf{s})$ become more symmetric as τ^2 increases.

Three types of predictor were considered, namely, ordinary kriging, ordinary constrained kriging, and ordinary covariance-matching constrained kriging. In the latter case, we actually considered a “partitioned” predictor corresponding to the partitioning of \mathbf{S} into subvectors of length five (i.e., $\mathbf{S} \equiv [S(1,1), \dots, S(1,5)|S(1,6), \dots, S(1,10)|\dots|S(10,6), \dots, S(10,10)]'$), except when the range $a = 4.1$, since no “partitioned” covariance-matching constrained kriging predictor exists for that partition of \mathbf{S} . In the case where $a = 4.1$, the predictor was formed by partitioning \mathbf{S} into pairs. The matrices \mathbf{P}_o and \mathbf{Q}_o , as defined in (5.22), were decomposed into symmetric components according to the spectral decomposition method.

6.3 Predictands

The predictands under consideration fall into three broad classes. In the first class, the following predictands were included,

$$\{S(D), S(B_1), S(B_2)\},$$

where

$$S(B) \equiv |B|^{-1} \sum_{s \in B} S(s); \quad B \subset D.$$

Note that these three predictands are linear functionals of $S(\cdot)$, and ordinary kriging is expected to predict these well, particularly if $S(\cdot)$ is Gaussian.

The second class of predictands consisted of

$$\{F_S(q(\alpha)) : \alpha = 0.75, 0.90\},$$

where

$$F_S(t) \equiv |D|^{-1} \sum_{s \in D} I(S(s) \leq t); \quad t \in \mathbb{R},$$

and $q(\alpha) \equiv \inf\{t : F_S(t) \geq \alpha\}$. Observe that $F_S(t)$ is a linear functional of univariate nonlinear functionals of $S(\cdot)$, and constrained kriging is expected to perform well here, particularly if $S(\cdot)$ is Gaussian.

Nonlinear functions of univariate nonlinear functionals of $S(\cdot)$ make up the third class of predictands, and this class includes

$$\{Q_A(q(\alpha)) : \alpha = 0.75, 0.90\},$$

where

$$Q_A(t) \equiv \sum_{s \in D} S(s) I(S(s) > t) / |D|(1 - F_S(t)); \quad t \in \mathbb{R},$$

and

$$\{Q_A^c(q(\alpha)) : \alpha = 0.75, 0.90\},$$

where

$$Q_A^c(t) \equiv \sum_{s \in D} S(s) I(S(s) \leq t) / |D|F_S(t); \quad t \in \mathbb{R}.$$

It was largely for predictors in this class that covariance-matching constrained kriging was developed.

6.4 Responses of the Experiment

We considered three responses in this experiment: The mean-squared prediction error (MSPE), the bias, and the variance of the predictors. Suppose that $\hat{p}(g(S(\cdot)); \mathbf{Z})$ is a predictor of the generic predictand $g(S(\cdot))$. Then the MSPE of $\hat{p}(g(S(\cdot)); \mathbf{Z})$ is

$$\begin{aligned} \text{MSPE}[\hat{p}(g(S(\cdot)); \mathbf{Z})] &\equiv E(\hat{p}(g(S(\cdot)); \mathbf{Z}) - g(S(\cdot)))^2 \\ &= \text{var}(\hat{p}(g(S(\cdot)); \mathbf{Z}) - g(S(\cdot))) + [E(\hat{p}(g(S(\cdot)); \mathbf{Z}) - g(S(\cdot)))]^2. \end{aligned}$$

The MSPE, bias, and variance of each predictor for all combinations of the factors and predictands of the experiment were estimated by generating 10,000 realizations of the appropriate S -processes and ϵ -processes, yielding

$$\begin{aligned} \widehat{\text{MSPE}}[\hat{p}(g(S(\cdot)); \mathbf{Z})] &\equiv \frac{1}{10000} \sum_{i=1}^{10000} (\hat{p}(g(S^{(i)}(\cdot)); \mathbf{Z}^{(i)}) - g(S^{(i)}(\cdot)))^2, \\ \widehat{\text{bias}}[\hat{p}(g(S(\cdot)); \mathbf{Z})] &\equiv \frac{1}{10000} \sum_{i=1}^{10000} (\hat{p}(g(S^{(i)}(\cdot)); \mathbf{Z}^{(i)}) - g(S^{(i)}(\cdot))), \end{aligned}$$

and

$$\widehat{\text{var}}[\hat{p}(g(S(\cdot)); \mathbf{Z})] \equiv \widehat{\text{MSPE}}[\hat{p}(g(S(\cdot)); \mathbf{Z})] - (\widehat{\text{bias}}[\hat{p}(g(S(\cdot)); \mathbf{Z})])^2,$$

where $S^{(i)}(\cdot)$ represents the i th simulation of the random field $S(\cdot)$; $i = 1, \dots, 10000$, $\mathbf{Z}^{(i)} \equiv (S^{(i)}(\mathbf{s}_1) + \epsilon^{(i)}(\mathbf{s}_1), \dots, S^{(i)}(\mathbf{s}_{12}) + \epsilon^{(i)}(\mathbf{s}_{12}))'$ represents the i th simulated data vector; $i = 1, \dots, 10000$, and $\epsilon^{(i)}(\cdot)$ represents the i th simulation of the measurement-error process; $i = 1, \dots, 10000$.

6.5 Results of the Experiment

In this computer-simulation experiment, all the covariance parameters of $Z(\cdot)$ (i.e., this includes the parameter τ^2) were completely specified, thus excluding the variability due to covariance-parameter estimation that usually occurs in practice. However, an advantage of specifying these parameters is that a “clean” comparison can be made among the predictors in question, without any confounding effects that may be due to their estimation. Moreover, estimated covariance parameters are used in the same “plug-in” fashion for each of the three predictors so that their effect is expected to be similar in each case.

Consider the linear predictands $\{S(D), S(B_1), S(B_2)\}$. All predictors of $S(D)$ performed comparably well, and this will not be discussed further here. Observe from Figure 6.1 that B_1 lies where the sample is most concentrated and that B_2 lies where the sample is least concentrated. The MSPE and squared bias for predictors of $S(B_1)$ and $S(B_2)$ are plotted in Figure 6.3 and Figure 6.4, respectively, and the MSPE, variance, and bias of those predictors are given in Table 6.2. In predicting $S(B_1)$ and $S(B_2)$, all predictors exhibited extremely low bias, irrespective of random field, range, or noise level. This is to be expected, since all predictors are unbiased. Not surprisingly, ordinary kriging exhibited the smallest variance and the smallest MSPE, in almost every case. However, covariance-matching constrained kriging consistently yielded a smaller variance and MSPE than constrained kriging, even though the former predictor is optimized with more constraints than the latter. In all cases, the MSPE was smallest when the random field was Gaussian, and all predictors performed similarly at the strongest level of spatial correlation. The MSPEs of predictors of $S(B_2)$ were consistently larger than those of $S(B_1)$.

The MSPE and squared bias for predictors of $F(q(0.75))$ and $F(q(0.90))$ are plotted in Figure 6.5 and Figure 6.6, respectively, and the MSPE, variance, and bias of those predictors are given in Table 6.3. For the prediction of $F(q(0.75))$, ordinary kriging consistently did worse than the other predictors, with respect to bias and MSPE, especially if the spatial correlation was weak. For the Gaussian random field $S_1(\cdot)$, both constrained kriging and covariance-matching constrained kriging exhibited low bias, with the latter predictor usually outperforming slightly the former, with respect to the bias criterion. For the two non Gaussian random fields, $S_2(\cdot)$ and $S_3(\cdot)$, ordinary kriging often yielded a smaller bias, although larger MSPE, than the other predictors. For the prediction of $F(q(0.90))$, ordinary kriging consistently performed comparably or even better than the other predictors, with respect to the MSPE, particularly when the range α was small. For the random field $S_1(\cdot)$, ordinary kriging consistently displayed a larger bias and a smaller variance than those of the other predictors, but for $S_2(\cdot)$ and $S_3(\cdot)$, the opposite was true.

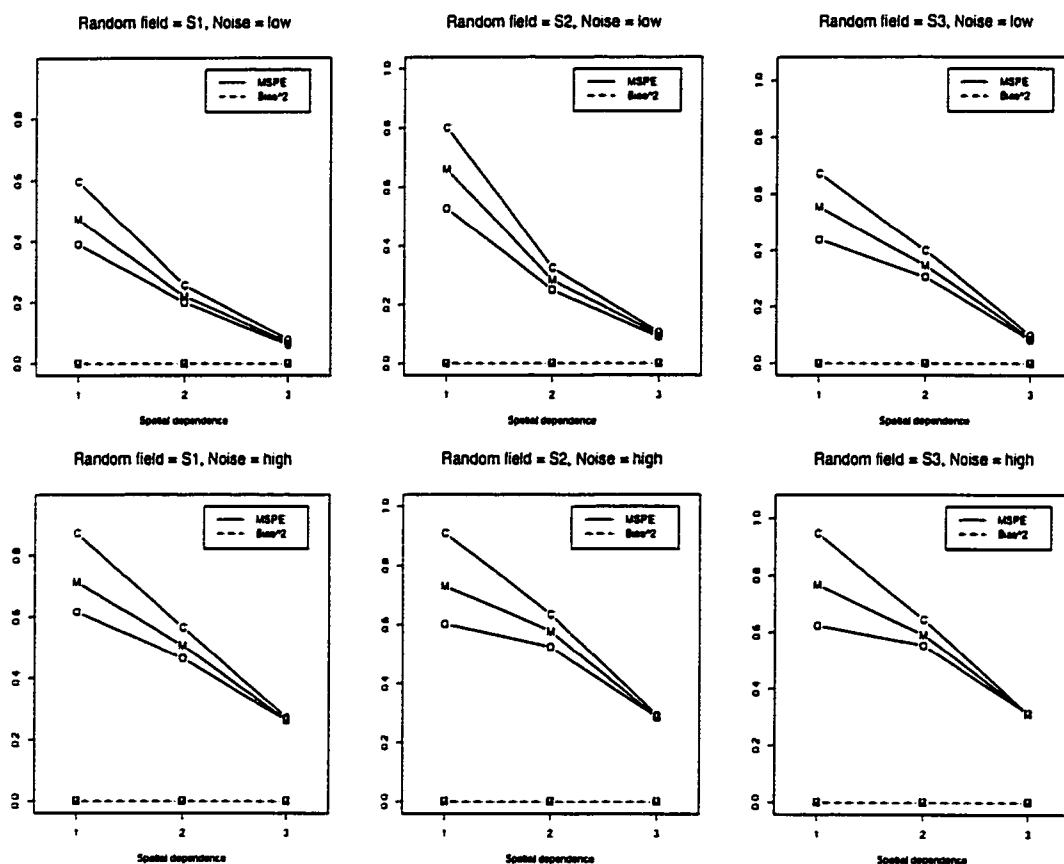


Figure 6.3 The MSPE and squared bias of predictors of $S(B_1)$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

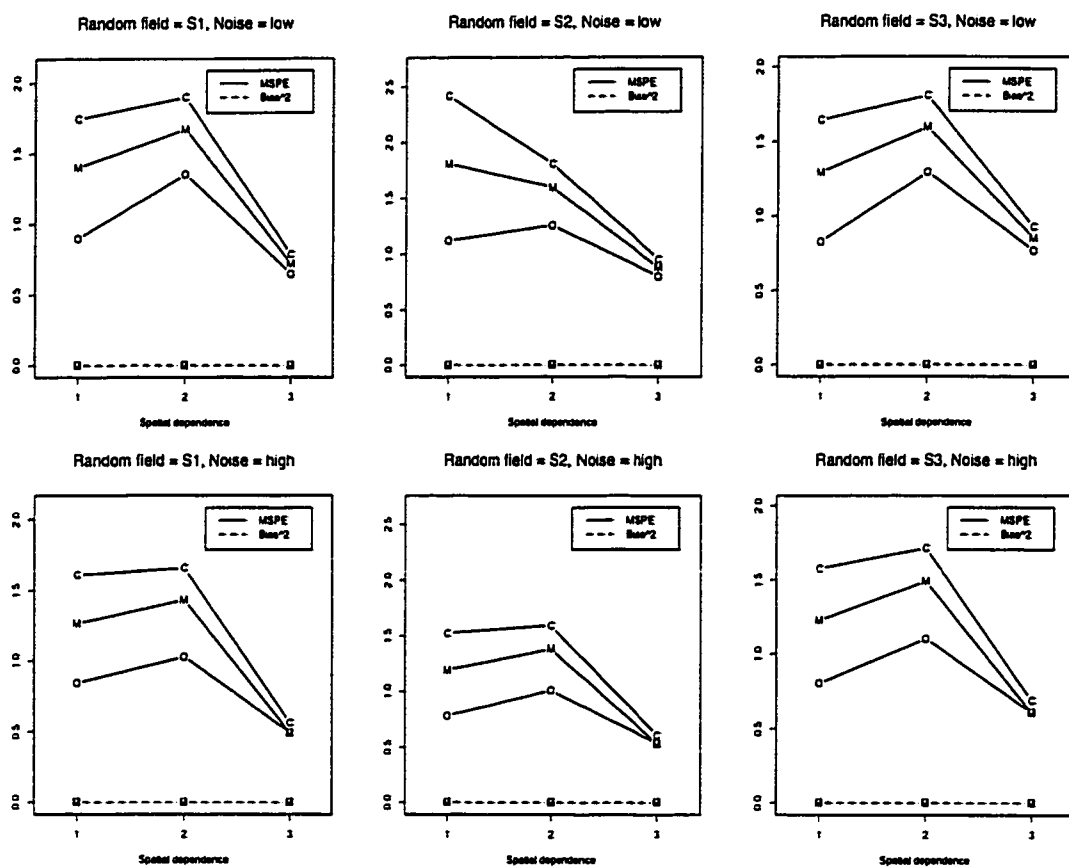


Figure 6.4 The MSPE and squared bias of predictors of $S(B_2)$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

Table 6.2 The MSPE, bias, and variance of predictors of $S(B_1)$ and $S(B_2)$, for all combinations of the experimental factors. Ordinary kriging is denoted "OK", constrained kriging "CK", and covariance-matching constrained kriging "CM".

$S(B_1)$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$a = 4.1$ $\tau^2 = 0.1$	OK	0.390	-0.002	0.390	0.528	0.009	0.528	0.437	-0.008	0.437
	CK	0.596	-0.005	0.596	0.800	0.001	0.800	0.672	-0.018	0.671
	CM	0.474	-0.003	0.474	0.660	0.004	0.660	0.554	-0.013	0.554
$a = 7.5$ $\tau^2 = 0.1$	OK	0.199	-0.003	0.199	0.248	-0.006	0.248	0.304	-0.012	0.304
	CK	0.256	-0.006	0.256	0.324	-0.003	0.324	0.398	-0.012	0.399
	CM	0.219	-0.005	0.219	0.282	-0.003	0.282	0.345	-0.012	0.344
$a = 30$ $\tau^2 = 0.1$	OK	0.062	0.002	0.062	0.090	-0.003	0.090	0.083	0.001	0.083
	CK	0.078	0.003	0.078	0.105	-0.004	0.105	0.098	0.004	0.098
	CM	0.069	0.002	0.069	0.097	-0.004	0.097	0.089	0.003	0.089
$a = 4.1$ $\tau^2 = 2$	OK	0.617	-0.005	0.617	0.602	0.009	0.602	0.623	0.001	0.623
	CK	0.872	-0.003	0.872	0.909	0.017	0.908	0.947	0.009	0.947
	CM	0.713	-0.005	0.713	0.729	0.014	0.729	0.765	0.005	0.765
$a = 7.5$ $\tau^2 = 2$	OK	0.465	0.007	0.465	0.524	-0.008	0.524	0.553	-0.004	0.553
	CK	0.565	0.004	0.565	0.634	-0.009	0.634	0.644	0.002	0.644
	CM	0.506	0.004	0.506	0.577	-0.008	0.576	0.591	-0.000	0.591
$a = 30$ $\tau^2 = 2$	OK	0.262	-0.001	0.262	0.285	-0.008	0.285	0.312	-0.006	0.312
	CK	0.270	-0.001	0.270	0.293	-0.008	0.293	0.312	-0.005	0.312
	CM	0.263	-0.001	0.263	0.286	-0.008	0.286	0.312	-0.006	0.312
$S(B_2)$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$a = 4.1$ $\tau^2 = 0.1$	OK	0.899	-0.005	0.899	1.121	-0.008	1.121	0.828	0.008	0.828
	CK	1.745	-0.004	1.745	2.417	-0.016	2.416	1.645	0.004	1.645
	CM	1.402	-0.005	1.402	1.810	-0.012	1.810	1.295	0.004	1.295
$a = 7.5$ $\tau^2 = 0.1$	OK	1.353	-0.002	1.353	1.256	-0.007	1.256	1.296	0.010	1.295
	CK	1.898	0.003	1.898	1.808	-0.010	1.808	1.809	0.012	1.808
	CM	1.673	0.001	1.673	1.597	-0.009	1.597	1.598	0.012	1.598
$a = 30$ $\tau^2 = 0.1$	OK	0.648	0.001	0.648	0.795	-0.000	0.795	0.769	-0.009	0.769
	CK	0.784	-0.000	0.784	0.951	0.000	0.951	0.926	-0.010	0.926
	CM	0.719	0.000	0.719	0.877	0.000	0.877	0.852	-0.009	0.852
$a = 4.1$ $\tau^2 = 2$	OK	0.846	0.004	0.846	0.783	0.003	0.783	0.805	0.009	0.804
	CK	1.611	0.003	1.611	1.525	0.007	1.524	1.577	0.014	1.576
	CM	1.266	0.005	1.266	1.193	0.005	1.193	1.227	0.011	1.227
$a = 7.5$ $\tau^2 = 2$	OK	1.031	-0.003	1.031	1.009	0.002	1.009	1.104	-0.004	1.104
	CK	1.658	-0.002	1.658	1.592	0.003	1.592	1.714	-0.003	1.714
	CM	1.432	-0.002	1.432	1.382	0.003	1.382	1.492	-0.003	1.492
$a = 30$ $\tau^2 = 2$	OK	0.493	0.004	0.493	0.532	-0.011	0.532	0.609	-0.005	0.609
	CK	0.564	0.003	0.564	0.606	-0.012	0.606	0.691	-0.005	0.691
	CM	0.493	0.004	0.493	0.532	-0.011	0.532	0.608	-0.004	0.608

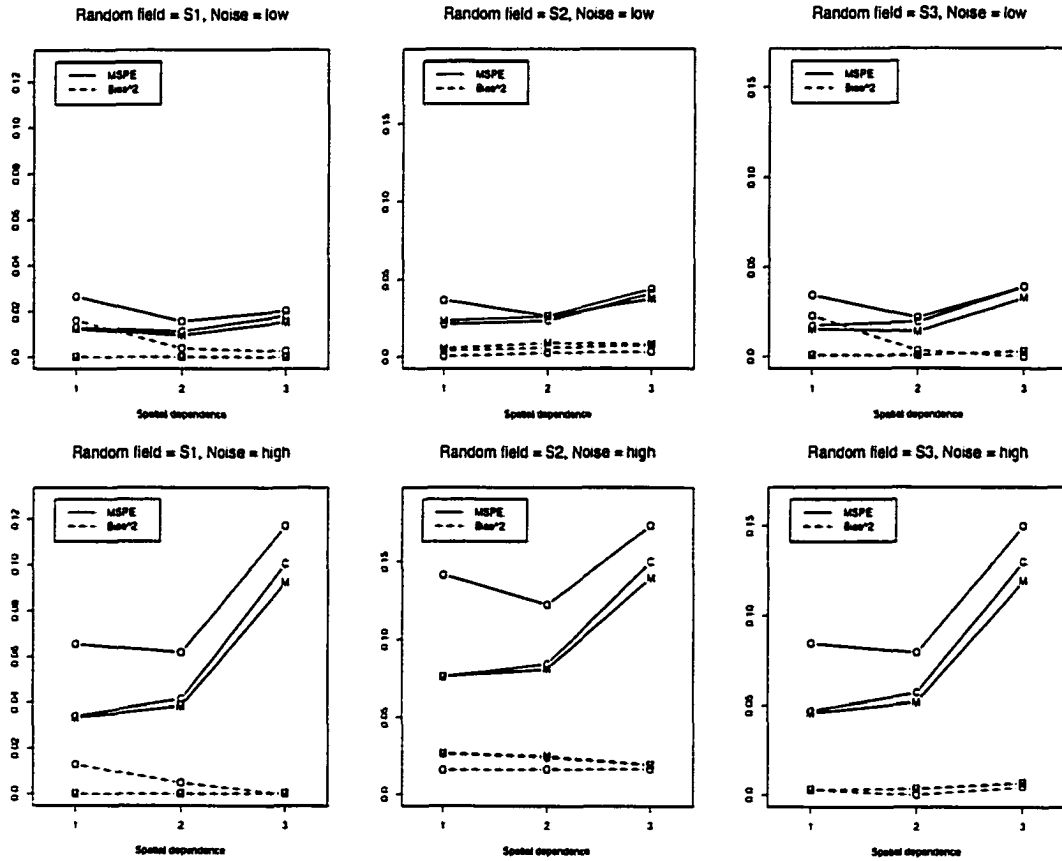


Figure 6.5 The MSPE and squared bias of predictors of $F_S(q(0.75))$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

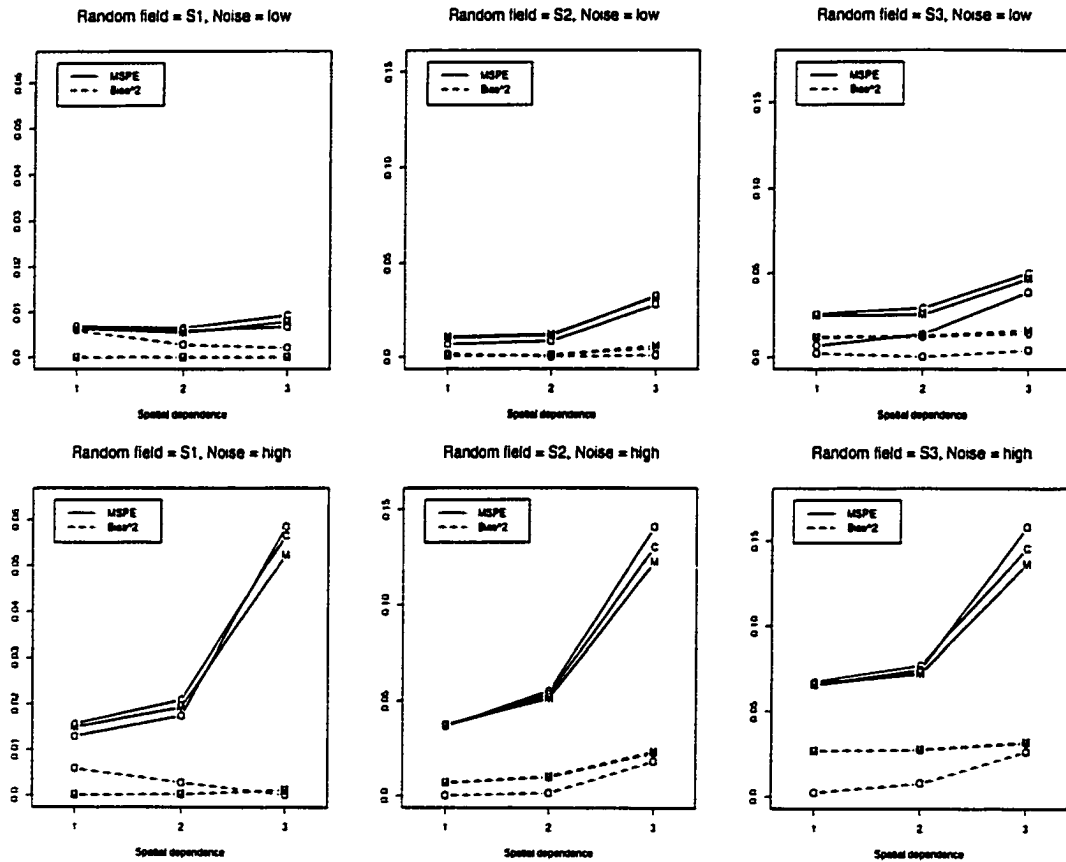


Figure 6.6 The MSPE and squared bias of predictors of $F_S(q(0.90))$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

Table 6.3 The MSPE, bias, and variance of predictors of $F_S(q(0.75))$ and $F_S(q(0.90))$, for all combinations of the experimental factors. Ordinary kriging is denoted "OK", constrained kriging "CK", and covariance-matching constrained kriging "CM".

$F(q(0.75))$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$a = 4.1$ $\tau^2 = 0.1$	OK	0.027	-0.126	0.011	0.037	0.020	0.036	0.034	-0.151	0.012
	CK	0.013	0.007	0.013	0.021	0.062	0.017	0.017	0.029	0.016
	CM	0.012	0.005	0.012	0.023	0.073	0.018	0.015	0.027	0.014
$a = 7.5$ $\tau^2 = 0.1$	OK	0.016	-0.061	0.012	0.026	0.051	0.023	0.022	-0.059	0.018
	CK	0.011	0.006	0.011	0.023	0.077	0.017	0.019	0.031	0.018
	CM	0.010	0.007	0.010	0.026	0.094	0.017	0.014	0.027	0.013
$a = 30$ $\tau^2 = 0.1$	OK	0.020	-0.051	0.018	0.043	0.056	0.040	0.039	-0.000	0.039
	CK	0.018	0.005	0.018	0.041	0.085	0.034	0.039	0.053	0.036
	CM	0.015	0.006	0.015	0.037	0.090	0.029	0.033	0.057	0.030
$a = 4.1$ $\tau^2 = 2$	OK	0.066	-0.114	0.053	0.141	0.125	0.126	0.084	-0.054	0.082
	CK	0.034	0.006	0.034	0.076	0.162	0.050	0.047	0.055	0.044
	CM	0.033	0.005	0.033	0.076	0.163	0.049	0.046	0.054	0.043
$a = 7.5$ $\tau^2 = 2$	OK	0.062	-0.070	0.057	0.122	0.128	0.106	0.079	-0.012	0.079
	CK	0.041	0.010	0.041	0.084	0.154	0.060	0.057	0.059	0.054
	CM	0.038	0.010	0.038	0.081	0.158	0.056	0.052	0.060	0.048
$a = 30$ $\tau^2 = 2$	OK	0.117	-0.001	0.117	0.173	0.128	0.156	0.150	0.069	0.145
	CK	0.101	0.018	0.100	0.150	0.137	0.131	0.130	0.082	0.123
	CM	0.093	0.021	0.092	0.139	0.137	0.120	0.119	0.083	0.112
$F(q(0.90))$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$a = 4.1$ $\tau^2 = 0.1$	OK	0.007	-0.077	0.001	0.007	-0.043	0.005	0.007	-0.049	0.004
	CK	0.007	0.006	0.007	0.010	0.022	0.010	0.026	0.107	0.014
	CM	0.006	0.005	0.006	0.011	0.024	0.010	0.025	0.110	0.013
$a = 7.5$ $\tau^2 = 0.1$	OK	0.006	-0.053	0.003	0.008	-0.012	0.008	0.014	0.018	0.013
	CK	0.006	0.005	0.006	0.011	0.029	0.010	0.029	0.110	0.017
	CM	0.005	0.007	0.005	0.012	0.034	0.011	0.025	0.114	0.012
$a = 30$ $\tau^2 = 0.1$	OK	0.007	-0.045	0.005	0.028	0.033	0.027	0.039	0.064	0.035
	CK	0.009	0.001	0.009	0.033	0.068	0.028	0.050	0.119	0.036
	CM	0.008	0.006	0.008	0.032	0.078	0.026	0.047	0.126	0.031
$a = 4.1$ $\tau^2 = 2$	OK	0.013	-0.076	0.007	0.036	-0.011	0.036	0.066	0.049	0.063
	CK	0.016	0.009	0.016	0.037	0.082	0.030	0.067	0.164	0.040
	CM	0.015	0.007	0.015	0.037	0.082	0.030	0.066	0.164	0.039
$a = 7.5$ $\tau^2 = 2$	OK	0.017	-0.052	0.015	0.055	0.038	0.054	0.074	0.088	0.066
	CK	0.021	0.015	0.021	0.053	0.098	0.044	0.077	0.165	0.050
	CM	0.019	0.016	0.019	0.051	0.101	0.041	0.072	0.166	0.045
$a = 30$ $\tau^2 = 2$	OK	0.058	0.011	0.058	0.141	0.135	0.123	0.158	0.162	0.131
	CK	0.056	0.032	0.056	0.130	0.150	0.108	0.145	0.177	0.114
	CM	0.052	0.035	0.051	0.123	0.152	0.099	0.136	0.179	0.104

Consider predicting $\{Q_A(q(\alpha)), Q_A^c(q(\alpha)) : \alpha = 0.75, 0.90\}$. Observe that $Q_A(t)$ is not defined if $1 - F_S(t) = 0$, and that $Q_A^c(t)$ is not defined if $F_S(t) = 0$. For $t = q(0.75)$ and $q(0.90)$, there were many occasions among the 10,000 simulations generated where $1 - F_S(t) = 0$. The proportion of all simulations for which $1 - F_S(t) > 0$ (i.e., the proportion for which $Q_A(t)$ is defined) is displayed in Figure 6.7, Figure 6.8, and Table 6.4. Covariance-matching constrained kriging has the highest proportion in all cases, and ordinary kriging has the lowest in all cases, sometimes substantially so. There were no cases where $F_S(t) = 0$; $t = q(0.75), q(0.90)$, and hence $Q_A^c(t)$ was always defined.

Table 6.4 The proportion of 10,000 simulations for which $1 - F_S(t) > 0$; $t = q(0.75), q(0.90)$, for all combinations of the experimental factors. Ordinary kriging is denoted "OK", constrained kriging "CK", and covariance-matching constrained kriging "CM".

Proportion $1 - F_S(t) > 0$							
		$t = q(0.75)$			$t = q(0.90)$		
		RF = $S_1(\cdot)$	RF = $S_2(\cdot)$	RF = $S_3(\cdot)$	RF = $S_1(\cdot)$	RF = $S_2(\cdot)$	RF = $S_3(\cdot)$
$a = 4.1$ $\tau^2 = 0.1$	OK	0.968	0.979	0.957	0.709	0.750	0.868
	CK	0.992	0.989	0.996	0.862	0.815	0.979
	CM	0.994	0.992	0.998	0.894	0.828	0.988
$a = 7.5$ $\tau^2 = 0.1$	OK	0.972	0.986	0.964	0.722	0.786	0.884
	CK	0.992	0.993	0.991	0.867	0.846	0.964
	CM	0.998	0.997	0.999	0.936	0.882	0.992
$a = 30$ $\tau^2 = 0.1$	OK	0.939	0.966	0.945	0.651	0.764	0.821
	CK	0.969	0.982	0.973	0.774	0.830	0.895
	CM	0.990	0.994	0.992	0.884	0.895	0.965
$a = 4.1$ $\tau^2 = 2$	OK	0.691	0.894	0.736	0.304	0.530	0.658
	CK	0.927	0.978	0.943	0.709	0.804	0.918
	CM	0.942	0.984	0.959	0.747	0.829	0.938
$a = 7.5$ $\tau^2 = 2$	OK	0.673	0.859	0.718	0.329	0.530	0.639
	CK	0.882	0.957	0.909	0.643	0.752	0.866
	CM	0.935	0.981	0.960	0.748	0.816	0.934
$a = 30$ $\tau^2 = 2$	OK	0.523	0.659	0.587	0.305	0.466	0.503
	CK	0.604	0.724	0.663	0.384	0.532	0.576
	CM	0.725	0.818	0.765	0.500	0.634	0.686

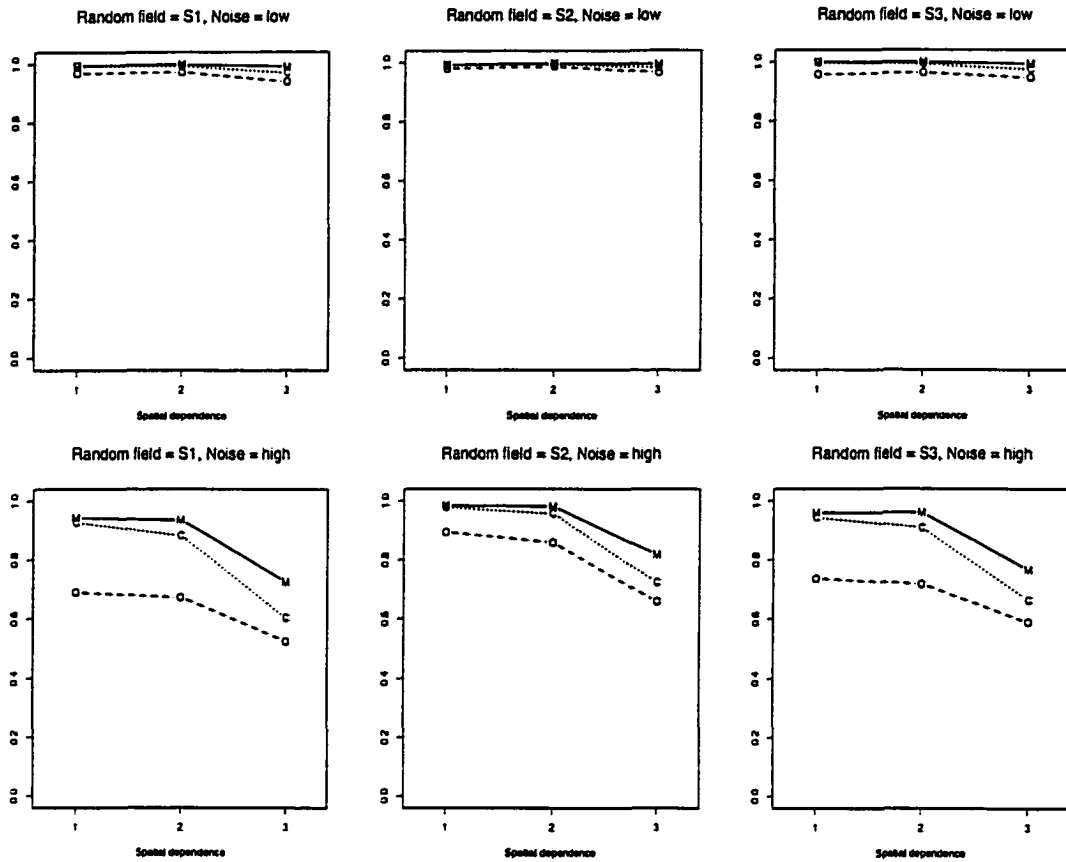


Figure 6.7 The proportion of 10,000 simulations for which $1 - F_S(t) > 0$ (i.e., the proportion for which $Q_A(q(0.75))$ is defined), for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1-3). In the plots, the lines marked "O" represent ordinary kriging, those marked "C" represent constrained kriging, and those marked "M" represent covariance-matching constrained kriging.

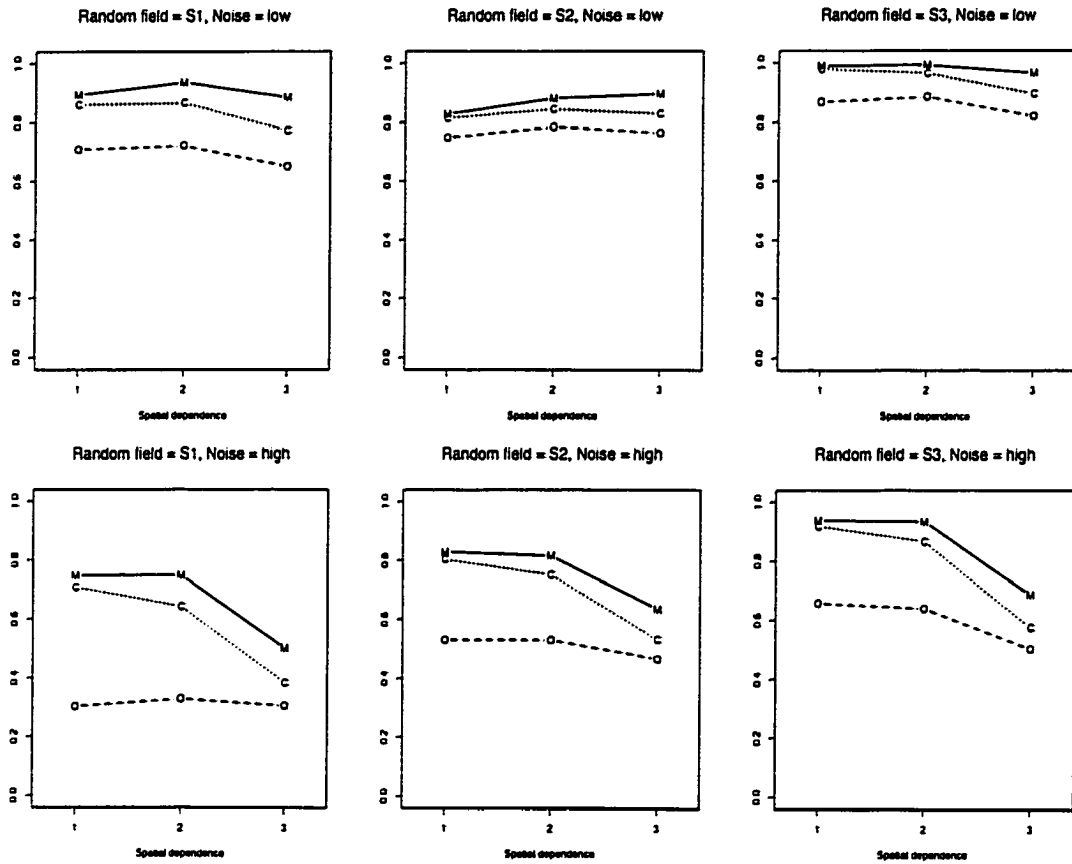


Figure 6.8 The proportion of 10,000 simulations for which $1 - F_S(t) > 0$ (i.e., the proportion for which $Q_A(q(0.90))$ is defined), for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3). In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

The MSPE and squared bias for predictors of $Q_A(q(0.75))$ and $Q_A(q(0.90))$ are plotted in Figure 6.9 and Figure 6.10, respectively, and the MSPE, variance, and bias of those predictors are given in Table 6.5. The performance of predictors of $Q_A(q(0.75))$ conditional on $1 - F_S(q(0.75)) > 0$, and $Q_A(q(0.90))$ conditional on $1 - F_S(q(0.90)) > 0$. That is, the MSPE is an optimistic estimate of the expected “loss” of the predictor if the corresponding proportion in Table 6.4 is less than one. Perhaps it would be more realistic to construct some loss function that takes into account both the MSPE and the proportion for which $Q_A(t)$; $t = q(0.75), q(0.90)$, is defined. In this study, we only compare the MSPE, bias, and variance of the predictors, but we must be mindful that these quantities are most optimistic for ordinary kriging since it consistently recorded the smallest proportion of simulations for which $Q_A(t)$ was defined, and least optimistic for covariance-matching constrained kriging since it consistently recorded the largest proportion of simulations for which $Q_A(t)$ was defined. In the Gaussian case, covariance-matching constrained kriging betters the other predictors, and ordinary kriging performs substantially worse unless the range is large. In the log Gaussian case, covariance-matching constrained kriging also betters the other predictors, although ordinary kriging does not perform as poorly as in the Gaussian case when the noise level is low, and constrained kriging performs as well as covariance-matching constrained kriging when the noise level is high. In the “mirror image” log Gaussian case, ordinary kriging substantially outperforms the other predictors, and covariance-matching constrained kriging performs worst especially when the noise level is low, unless the range is large. Ordinary kriging generally has the largest bias if the random field is $S_1(\cdot)$ or $S_2(\cdot)$, but the smallest bias for $S_3(\cdot)$. Differences in performance among the predictors is more pronounced in the case $t = q(0.75)$.

The MSPE and squared bias for predictors of $Q_A^c(q(0.75))$ and $Q_A^c(q(0.90))$ are plotted in Figure 6.11 and Figure 6.12, respectively, and the MSPE, variance, and bias of those predictors are given in Table 6.6. Predictors of $Q_A^c(t)$ perform somewhat similarly to those of $Q_A(t)$ although, in predicting the former, ordinary kriging does better in the log Gaussian case with respect to the MSPE, and covariance-matching constrained kriging does better if the random field is $S_3(\cdot)$ and $\tau^2 = 0.1$.

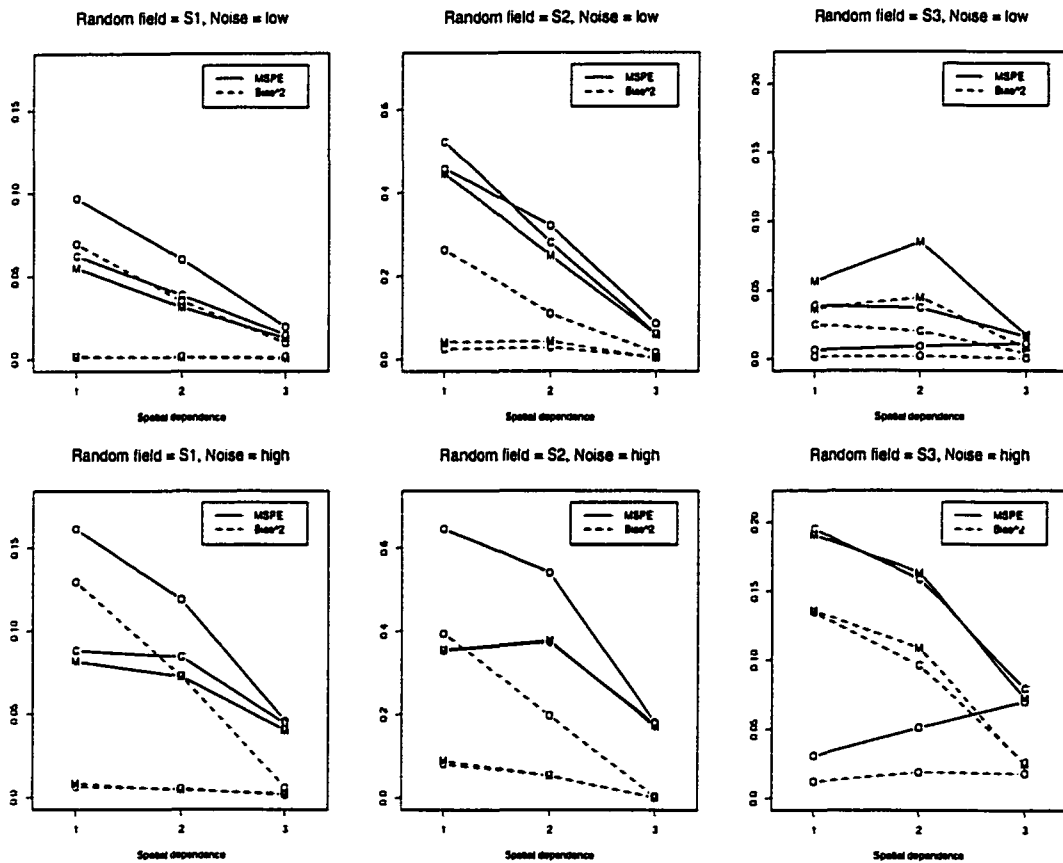


Figure 6.9 The MSPE and squared bias of predictors of $Q_A(q(0.75))$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

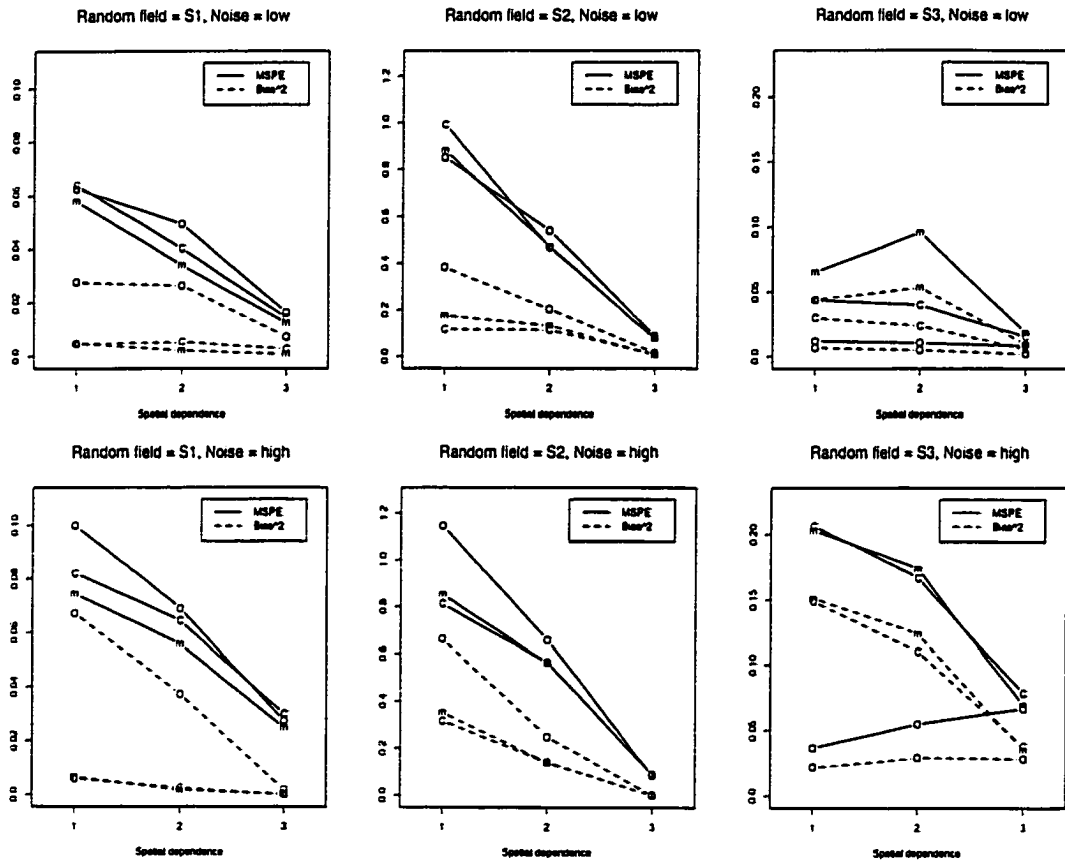


Figure 6.10 The MSPE and squared bias of predictors of $Q_A(q(0.90))$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

Table 6.5 The MSPE, bias, and variance of predictors of $Q_A(q(0.75))$ conditional on $1 - F_S(q(0.75)) > 0$, and $Q_A(q(0.90))$ conditional on $1 - F_S(q(0.90)) > 0$, for all combinations of the experimental factors. Ordinary kriging is denoted "OK", constrained kriging "CK", and covariance-matching constrained kriging "CM".

$Q_A(q(0.75))$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$a = 4.1$ $\tau^2 = 0.1$	OK	0.097	0.264	0.027	0.458	0.513	0.195	0.006	-0.042	0.005
	CK	0.062	0.037	0.061	0.522	0.157	0.497	0.039	-0.158	0.014
	CM	0.055	0.044	0.053	0.445	0.205	0.403	0.056	-0.191	0.020
$a = 7.5$ $\tau^2 = 0.1$	OK	0.060	0.187	0.025	0.320	0.332	0.210	0.009	-0.046	0.007
	CK	0.039	0.041	0.037	0.280	0.170	0.251	0.037	-0.142	0.017
	CM	0.031	0.031	0.030	0.249	0.208	0.205	0.085	-0.211	0.040
$a = 30$ $\tau^2 = 0.1$	OK	0.020	0.100	0.010	0.086	0.125	0.070	0.011	-0.020	0.010
	CK	0.015	0.042	0.013	0.061	0.068	0.056	0.016	-0.064	0.012
	CM	0.013	0.025	0.012	0.059	0.054	0.056	0.017	-0.087	0.010
$a = 4.1$ $\tau^2 = 2$	OK	0.161	0.360	0.032	0.644	0.628	0.250	0.031	-0.110	0.019
	CK	0.088	0.081	0.082	0.352	0.286	0.271	0.195	-0.367	0.060
	CM	0.082	0.092	0.073	0.354	0.299	0.265	0.191	-0.368	0.055
$a = 7.5$ $\tau^2 = 2$	OK	0.119	0.269	0.046	0.540	0.443	0.343	0.051	-0.138	0.032
	CK	0.084	0.074	0.079	0.372	0.230	0.319	0.159	-0.310	0.063
	CM	0.073	0.068	0.068	0.377	0.232	0.323	0.163	-0.330	0.055
$a = 30$ $\tau^2 = 2$	OK	0.046	0.078	0.040	0.180	0.051	0.177	0.070	-0.134	0.052
	CK	0.044	0.042	0.043	0.177	0.024	0.177	0.079	-0.161	0.053
	CM	0.040	0.051	0.037	0.170	0.030	0.169	0.072	-0.155	0.048
$Q_A(q(0.90))$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$a = 4.1$ $\tau^2 = 0.1$	OK	0.063	0.167	0.035	0.850	0.618	0.469	0.012	-0.083	0.005
	CK	0.064	0.070	0.059	0.992	0.342	0.875	0.044	-0.173	0.014
	CM	0.058	0.071	0.053	0.880	0.420	0.704	0.065	-0.210	0.021
$a = 7.5$ $\tau^2 = 0.1$	OK	0.050	0.163	0.023	0.539	0.451	0.336	0.011	-0.072	0.005
	CK	0.040	0.074	0.035	0.467	0.340	0.352	0.040	-0.155	0.016
	CM	0.034	0.049	0.032	0.473	0.365	0.339	0.096	-0.231	0.043
$a = 30$ $\tau^2 = 0.1$	OK	0.016	0.086	0.009	0.089	0.124	0.073	0.008	-0.045	0.006
	CK	0.014	0.054	0.011	0.079	0.100	0.069	0.015	-0.075	0.009
	CM	0.012	0.031	0.012	0.076	0.083	0.069	0.018	-0.097	0.009
$a = 4.1$ $\tau^2 = 2$	OK	0.100	0.259	0.033	1.145	0.814	0.482	0.037	-0.148	0.015
	CK	0.082	0.077	0.076	0.812	0.562	0.496	0.206	-0.386	0.058
	CM	0.074	0.081	0.068	0.851	0.592	0.500	0.203	-0.388	0.052
$a = 7.5$ $\tau^2 = 2$	OK	0.069	0.193	0.032	0.659	0.498	0.412	0.055	-0.171	0.026
	CK	0.064	0.051	0.062	0.563	0.374	0.423	0.167	-0.332	0.057
	CM	0.056	0.042	0.054	0.561	0.373	0.422	0.174	-0.352	0.049
$a = 30$ $\tau^2 = 2$	OK	0.027	0.043	0.026	0.084	-0.016	0.084	0.067	-0.169	0.038
	CK	0.030	0.013	0.030	0.090	-0.037	0.088	0.079	-0.195	0.041
	CM	0.025	0.017	0.024	0.089	-0.026	0.089	0.070	-0.187	0.035

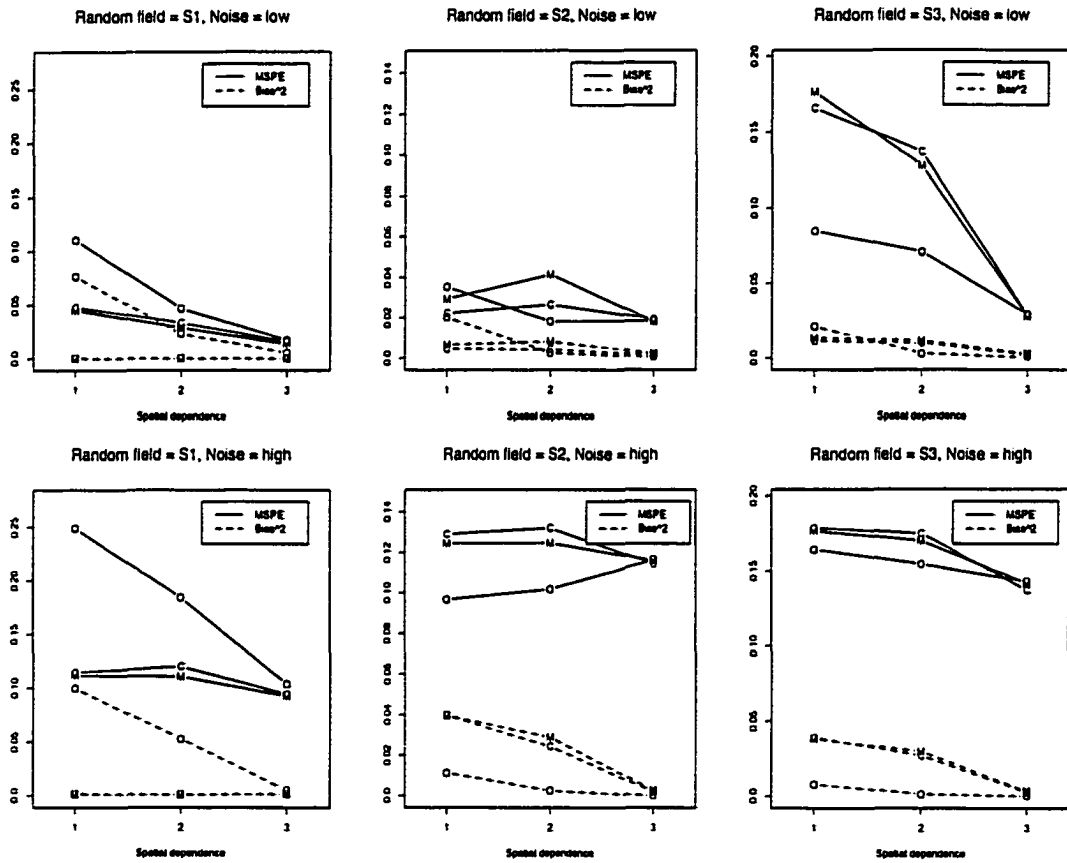


Figure 6.11 The MSPE and squared bias of predictors of $Q_A^c(q(0.75))$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

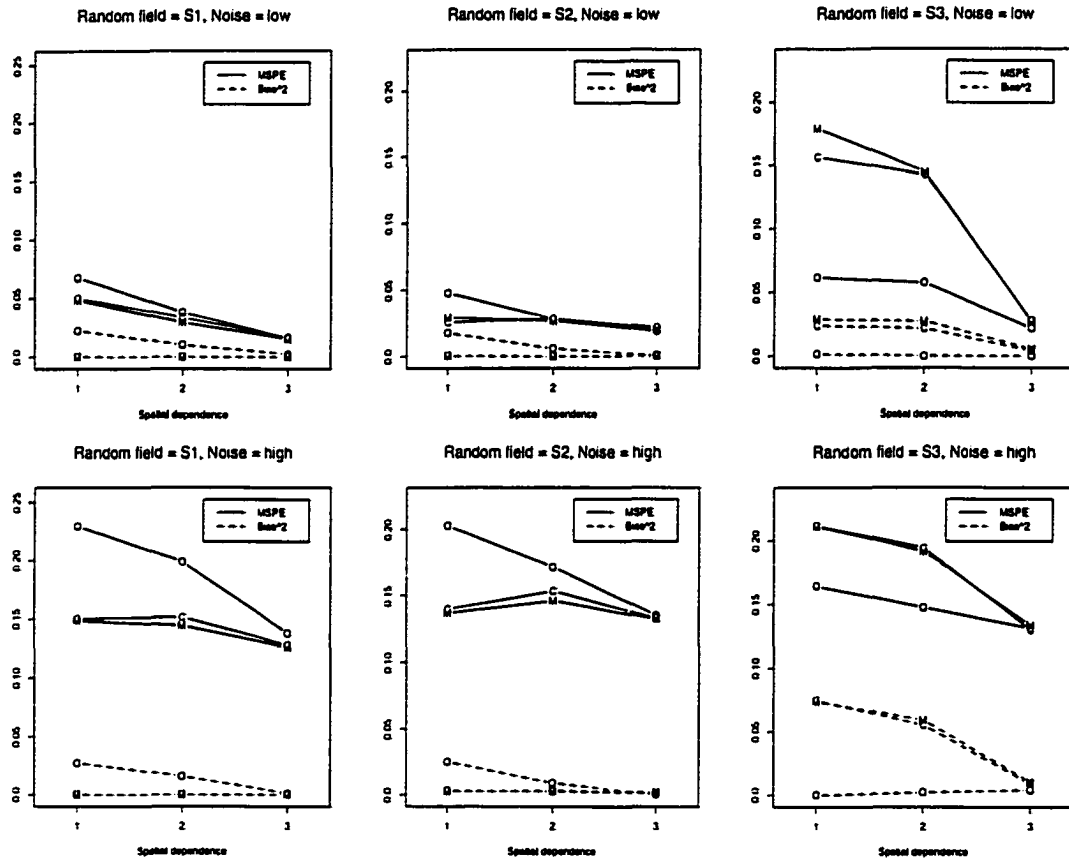


Figure 6.12 The MSPE and squared bias of predictors of $Q_A^c(q(0.90))$ are displayed here, for all combinations of the 3 random fields and the 2 noise levels. Along the horizontal axis of each plot, the three levels of the strength of spatial correlation are given in increasing order (marked 1–3), and the scale of the vertical axis varies with random field. In the plots, the lines marked “O” represent ordinary kriging, those marked “C” represent constrained kriging, and those marked “M” represent covariance-matching constrained kriging.

Table 6.6 The MSPE, bias, and variance of predictors of $Q_A^c(q(0.75))$ and $Q_A^c(q(0.90))$, for all combinations of the experimental factors. Ordinary kriging is denoted "OK", constrained kriging "CK", and covariance-matching constrained kriging "CM".

$Q_A^c(q(0.75))$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$\alpha = 4.1$ $\tau^2 = 0.1$	OK	0.110	-0.277	0.034	0.035	-0.142	0.015	0.085	-0.145	0.064
	CK	0.048	0.011	0.047	0.022	0.068	0.018	0.166	0.107	0.155
	CM	0.045	0.004	0.045	0.029	0.081	0.023	0.176	0.116	0.163
$\alpha = 7.5$ $\tau^2 = 0.1$	OK	0.046	-0.152	0.023	0.018	-0.050	0.016	0.071	-0.054	0.068
	CK	0.033	0.005	0.033	0.026	0.064	0.022	0.137	0.100	0.127
	CM	0.029	0.006	0.029	0.041	0.090	0.033	0.128	0.107	0.116
$\alpha = 30$ $\tau^2 = 0.1$	OK	0.017	-0.073	0.012	0.018	-0.021	0.018	0.029	-0.023	0.028
	CK	0.015	-0.000	0.015	0.019	0.039	0.018	0.028	0.044	0.026
	CM	0.013	0.002	0.013	0.018	0.048	0.016	0.027	0.052	0.025
$\alpha = 4.1$ $\tau^2 = 2$	OK	0.249	-0.316	0.149	0.097	-0.106	0.086	0.164	-0.088	0.157
	CK	0.114	-0.032	0.113	0.129	0.199	0.089	0.178	0.197	0.140
	CM	0.112	-0.038	0.110	0.124	0.199	0.085	0.177	0.195	0.139
$\alpha = 7.5$ $\tau^2 = 2$	OK	0.184	-0.230	0.131	0.102	-0.049	0.099	0.154	-0.038	0.153
	CK	0.120	-0.030	0.119	0.132	0.156	0.108	0.175	0.164	0.148
	CM	0.111	-0.029	0.110	0.125	0.170	0.096	0.170	0.172	0.140
$\alpha = 30$ $\tau^2 = 2$	OK	0.104	-0.070	0.099	0.116	0.011	0.116	0.143	0.018	0.143
	CK	0.094	-0.039	0.093	0.114	0.044	0.112	0.137	0.051	0.134
	CM	0.093	-0.040	0.091	0.116	0.050	0.114	0.142	0.057	0.139
$Q_A^c(q(0.90))$										
		Random Field = $S_1(\cdot)$			Random Field = $S_2(\cdot)$			Random Field = $S_3(\cdot)$		
		MSPE	Bias	Var	MSPE	Bias	Var	MSPE	Bias	Var
$\alpha = 4.1$ $\tau^2 = 0.1$	OK	0.067	-0.150	0.045	0.048	-0.133	0.030	0.062	-0.042	0.060
	CK	0.050	0.011	0.050	0.026	0.021	0.025	0.156	0.155	0.132
	CM	0.048	0.007	0.048	0.029	0.021	0.029	0.179	0.170	0.150
$\alpha = 7.5$ $\tau^2 = 0.1$	OK	0.038	-0.102	0.027	0.028	-0.077	0.022	0.058	0.013	0.057
	CK	0.033	0.005	0.033	0.028	0.009	0.028	0.142	0.147	0.121
	CM	0.030	0.007	0.030	0.027	0.010	0.027	0.144	0.166	0.117
$\alpha = 30$ $\tau^2 = 0.1$	OK	0.016	-0.049	0.014	0.019	-0.026	0.018	0.022	0.009	0.021
	CK	0.017	-0.002	0.017	0.022	0.014	0.022	0.028	0.063	0.024
	CM	0.015	0.004	0.015	0.020	0.023	0.019	0.027	0.073	0.022
$\alpha = 4.1$ $\tau^2 = 2$	OK	0.230	-0.165	0.202	0.203	-0.157	0.178	0.164	0.011	0.164
	CK	0.150	-0.010	0.150	0.139	0.054	0.136	0.212	0.273	0.137
	CM	0.148	-0.014	0.148	0.136	0.051	0.134	0.211	0.272	0.137
$\alpha = 7.5$ $\tau^2 = 2$	OK	0.199	-0.128	0.183	0.171	-0.095	0.162	0.147	0.049	0.145
	CK	0.152	-0.006	0.152	0.153	0.052	0.151	0.194	0.235	0.139
	CM	0.144	-0.004	0.144	0.146	0.057	0.143	0.191	0.243	0.132
$\alpha = 30$ $\tau^2 = 2$	OK	0.138	-0.034	0.136	0.135	0.014	0.135	0.131	0.066	0.127
	CK	0.128	-0.012	0.128	0.133	0.039	0.131	0.130	0.095	0.121
	CM	0.126	-0.012	0.126	0.132	0.046	0.130	0.133	0.102	0.123

6.6 Summary

Several general conclusions can be made from this computer-simulation experiment:

1. Ordinary kriging performs best for the prediction of linear predictands.
2. Covariance-matching constrained kriging outperforms constrained kriging if the random field representing the signal process is Gaussian.
3. Ordinary kriging performs best in predicting multivariate nonlinear predictands if the random field representing the signal process is left-skewed.
4. If the level of spatial correlation of the signal process is very strong, then the three predictors perform similarly, irrespective of predictand.

7 Spatial Analysis of Particulate Matter in Pittsburgh Area

In this section, we shall apply the linear kriging predictors encountered in Sections 3, 4, and 5 to the analysis of PM_{10} (i.e., particulate matter with particles smaller than 10 microns in aerodynamic diameter, in a cubic meter of air; see Section 1) over the Pittsburgh area.

In 1996, PM_{10} was monitored at 27 sites in the Pittsburgh area. The monitoring stations were scattered throughout the Pittsburgh Metropolitan Area and environs. At the 27 sites, four different monitoring devices (probe types) were used, at several sites more than one probe type was used, and at two sites the same probe type was duplicated. In addition, the monitoring intervals varied from daily to 3-day to 6-day (synchronous) intervals.

7.1 Exploratory Data Analysis

The exploratory data analysis consisted of two phases (Cressie et al., 1999). In the first phase, the analysis focused on differences among the probe types and appropriate transformations of the data to stabilize the variance and to obtain an approximately symmetric distribution; this analysis was based on all the data recorded in 1996.

Observations corresponding to the two most frequent probe types were compared (the other two probe types were too poorly represented for any meaningful comparisons), from which it was concluded that all probe types were essentially measuring the same quantity with the same measurement error, resulting in more sites with replicate observations. Various transformations from among the power family of transformations were considered, but a standard log transformation performed well in stabilizing the variance and producing a normal distribution in the data.

In the second phase, an exploratory *spatial* data analysis was performed on the data corresponding to a particular day, August 13, 1996 (day 226), upon which a detailed spatial analysis was done for illustrative purposes. Day 226 was chosen because a graphical analysis of the corresponding data exhibited a spatial-dependence structure sufficiently interesting to warrant a more detailed analysis. On this day, PM_{10} was monitored at 22 sites, eight of which had duplicate measurements. Figure 7.1 shows a map of the configuration of these 22 sites, centered near Avalon, about 5 miles north-west of downtown Pittsburgh, and extending about 20 miles in each of the compass directions. At the scale of the problem, the earth's surface is approximately flat and the units on each axis are given in miles.

An analysis of semivariogram clouds detected two outlying values (i.e., unusual values relative to nearby values) corresponding to a single site marked with an "x" in Figure 7.1. This site (site 3004) was removed for the rest of the exploratory spatial data analysis. Further analysis of semivariogram clouds and semivariogram estimates (with site 3004 removed) in several directions indicated that the spatial variability was *anisotropic* (i.e., not homogeneous in all directions), and that the spatial correlation appeared to be strongest in the ENE-WSW direction (i.e., 67.5° clockwise from N) and weakest in the NNE-SSW direction (i.e., 157.5° clockwise from N).

7.2 Modeling the Semivariogram

Define

$$S(s) \equiv \log PM_{10}(s); \quad s \in D,$$

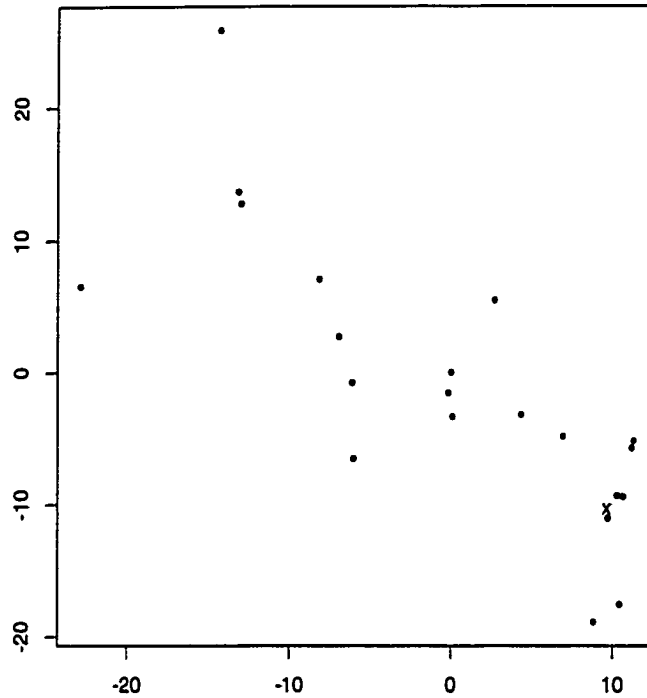


Figure 7.1 Map of configuration of 22 sites on day 226, 1996. The site marked "x" corresponds to the location of two spatial outliers. The horizontal axis is in the east-west direction, the vertical axis is in the north-south direction, and the units of both axes are in miles.

where D represents a two-dimensional convex region surrounding Pittsburgh and containing the 22 sampling locations. Assuming that the measurement error is normal and additive on the log-transformed scale, we obtain the measurement-error spatial model,

$$Z(s) \equiv S(s) + \epsilon(s); \quad s \in D,$$

where $Z(\cdot)$ represents the set of actual (and potential) observations on the log-transformed scale, and for each $s \in D$, $\epsilon(s)$ is distributed identically and independently as $N(0, \tau^2)$. In addition, $S(\cdot)$ and $\epsilon(\cdot)$ are assumed to be independent.

Nonparametric estimates of the semivariogram of the data (excluding site 3004) were cal-

culated (see Cressie, 1993a; Sect. 2.4), yielding

$$\{\hat{\gamma}_z(\mathbf{h}_i) : i = 1, \dots, H\},$$

where \mathbf{h}_i ; $i = 1, \dots, H$, are lag vectors taken in only two directions 67.5° and 157.5° , each with an angle tolerance of $\pm 45^\circ$ (i.e., if the direction of the line linking any pair of sites was within 45° of 67.5° , it was considered to be in the direction 67.5° ; otherwise it was considered to be in the direction 157.5°), and a lag tolerance of 1.5 miles (i.e., any distance between pairs that lay within ± 1.5 miles of $\|\mathbf{h}_i\|$ was considered to be of length $\|\mathbf{h}_i\|$; $i = 1, \dots, H$). The semivariogram estimates $\{\hat{\gamma}_z(\mathbf{h}_i)\}$ are displayed in Figure 7.2.

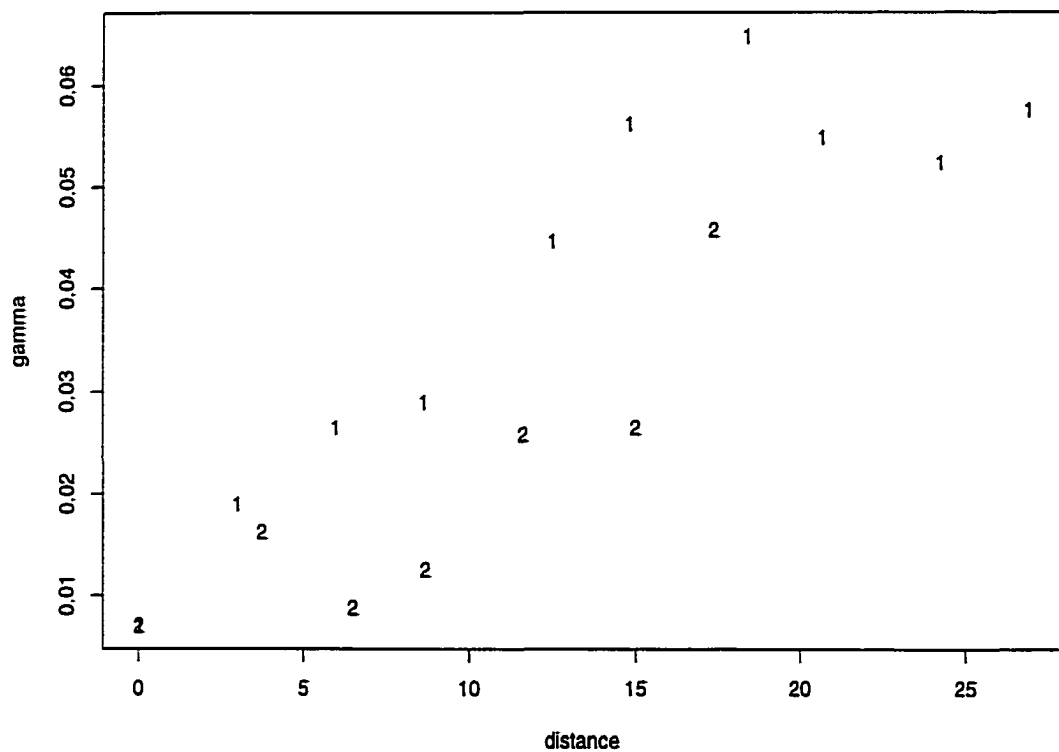


Figure 7.2 Semivariogram estimates are displayed in two directions: 157.5° (marked as “1”) and 67.5° (marked as “2”). An angle tolerance of 45° and a lag tolerance of 1.5 miles were used in both cases. The horizontal axis represents lag distance in miles, and the vertical axis represents estimated semivariogram values.

A semivariogram is said to possess *geometric anisotropy* if a simple rotation and scaling of the original axes exists such that under the transformed axes the semivariogram is isotropic (i.e., homogeneous in all directions; see Section 2 in the vicinity of (2.13)). Assuming geometric anisotropy for the directional semivariograms in Figure 7.2, Journel and Huijbregts (1978, pp. 179–181) give the matrix that performs the appropriate rotation and scaling of the original axes as

$$\mathbf{A} = \begin{bmatrix} \cos^2 \phi + \lambda \sin^2 \phi & (1 - \lambda) \sin \phi \cos \phi \\ (1 - \lambda) \sin \phi \cos \phi & \sin^2 \phi + \lambda \cos^2 \phi \end{bmatrix}, \quad (7.1)$$

where ϕ represents the angle of rotation (clockwise) required to align the original N-S axis with the direction in which the spatial correlation is weakest. Therefore, we obtain $\phi = 157.5^\circ$. The anisotropy-ratio parameter λ represents the scaling of the axes required for isotropy under the new axes, and it is estimated by weighted least squares below.

Applying the required rotation and scaling of the original axes, we obtain the *isotropic* semivariogram $\gamma_z^o(\cdot)$, where γ_z^o satisfies

$$\gamma_z^o(\|\mathbf{A}\mathbf{h}\|) \equiv \gamma_z(\mathbf{h}),$$

with $\mathbf{h} = \mathbf{s} - \mathbf{u}$; $\mathbf{s}, \mathbf{u} \in D$, and \mathbf{A} given by (7.1).

We are now ready to fit an isotropic semivariogram model by the method of weighted least squares, proposed by Cressie (1985), to the isotropic semivariogram estimates $\{\hat{\gamma}_z^o(\|\mathbf{A}\mathbf{h}_i\|) : i = 1, \dots, H\}$, where $\hat{\gamma}_z^o(\|\mathbf{A}\mathbf{h}_i\|) \equiv \hat{\gamma}_z(\mathbf{h}_i)$; $i = 1, \dots, H$. That is, we choose $\hat{\theta}$ in $\gamma_z^o(\|\mathbf{A}(\lambda)\mathbf{h}\|; \theta)$ to minimize

$$\sum_{i=1}^H N(\|\mathbf{A}(\lambda)\mathbf{h}_i\|) \left\{ \frac{\hat{\gamma}_z^o(\|\mathbf{A}(\lambda)\mathbf{h}_i\|)}{\gamma_z^o(\|\mathbf{A}(\lambda)\mathbf{h}\|; \theta)} - 1 \right\}^2,$$

with respect to $\theta \equiv (\lambda, \sigma_\eta^2, \sigma_W^2, a)'$, where we have chosen to feature the anisotropy-ratio parameter λ in $\mathbf{A}(\lambda)$, and $N(\|\mathbf{A}(\lambda)\mathbf{h}_i\|)$ is the number of pairs of sites whose distance (after the axes have been transformed by $\mathbf{A}(\lambda)$) lies within a tolerance of 1.5 miles of $\|\mathbf{A}(\lambda)\mathbf{h}_i\|$. The parameters σ_η^2 , σ_W^2 , and a are parameters of the spherical semivariogram model defined in Section 3.3. This results in the estimated (geometrically anisotropic) semivariogram model

given by

$$\gamma_z(\mathbf{h}; \hat{\theta}) = \begin{cases} \hat{\sigma}_\eta^2 + \hat{\sigma}_W^2, & \|\mathbf{A}(\hat{\lambda})\mathbf{h}\| > \hat{a}, \\ \hat{\sigma}_\eta^2 + \hat{\sigma}_W^2 \left\{ \frac{3}{2}(\|\mathbf{A}(\hat{\lambda})\mathbf{h}\|/\hat{a}) - \frac{1}{2}(\|\mathbf{A}(\hat{\lambda})\mathbf{h}\|/\hat{a})^3 \right\}, & 0 < \|\mathbf{A}(\hat{\lambda})\mathbf{h}\| \leq \hat{a}, \\ 0, & \mathbf{h} = \mathbf{0}, \end{cases} \quad (7.2)$$

where $\hat{\lambda} = 0.46923$, $\hat{\sigma}_\eta^2 = 0.00726$, $\hat{\sigma}_W^2 = 0.05188$, and $\hat{a} = 22.720$. The isotropic semivariogram is displayed in Figure 7.3.

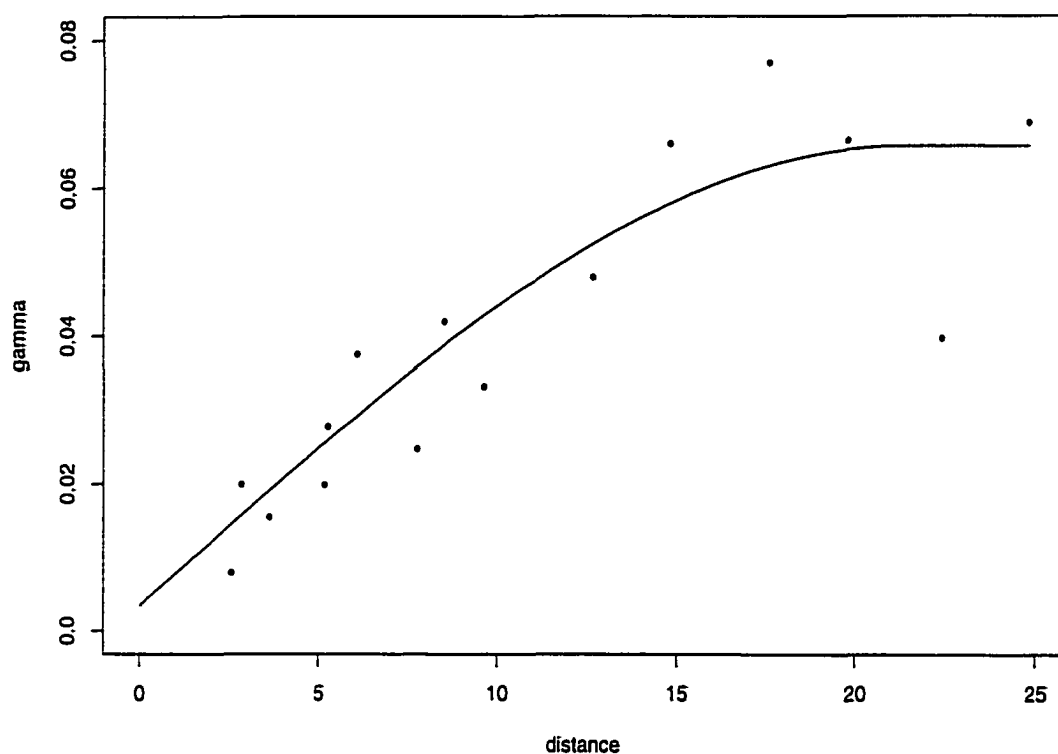


Figure 7.3 Fitted isotropic spherical semivariogram. The lag distance is the scalar quantity $\|\mathbf{A}\mathbf{h}\|$, where \mathbf{h} is the lag-distance vector between sites in the original co-ordinate system. The horizontal axis represents lag distance in miles, and the vertical axis represents fitted semivariogram values.

Finally, from the seven duplicated data, we can estimate the measurement-error variance τ^2 as $\hat{\tau}^2 = 0.00706$, which allows us to construct Σ and $c(B)$ for duplicated data. For example, $\text{var}(Z(s_i)) = \sigma^2 + \tau^2$, and $\text{cov}(Z_1(s_i), Z_2(s_i)) = \sigma^2$, where $Z_1(s_i)$ and $Z_2(s_i)$ represent duplicated data, and $\text{var}(S(s)) \equiv \sigma^2$.

7.3 Spatial Prediction of PM_{10}

Define

$$X(s) \equiv \text{PM}_{10}(s) = \exp(S(s)); \quad s \in D,$$

where D represents a two-dimensional convex region surrounding Pittsburgh and containing the 22 sampling locations.

In this section, we address several prediction objectives. A “smooth” surface of PM_{10} values is predicted to provide some idea of the nature of PM_{10} over D . Observed and predicted PM_{10} -values indicate that the region D is well within PM_{10} compliance (i.e., $\text{PM}_{10}(s) \leq 150 \mu\text{g}/\text{m}^3$; $s \in D$), so prediction of the proportion of D exceeding $150 \mu\text{g}/\text{m}^3$ is expected to be zero. However, for illustrative purposes, two arbitrary cutoff values (both within the range of observed values) were specified, and the proportion of D exceeding those two values is predicted. Finally, the average amount of PM_{10} that exceeds/does not exceed the two cutoff values are also predicted. The two outliers discussed in Section 7.1 were included in the prediction part of this study since, although it may be justifiable to exclude the strong local effects unusual observations may have on a global variogram model, that justification may not hold for the purposes of prediction, where local effects usually are of interest.

In predicting the map of PM_{10} values, D was discretized into a grid of 500 prediction locations, at each of which $X(s_0)$ was predicted by ordinary kriging and constrained kriging. Define the ordinary kriging predictor of $X(s_0)$ as

$$\hat{X}_{ok}(s_0) \equiv \exp(\hat{S}_{ok}(s_0)); \quad s_0 \in D,$$

where $\hat{S}_{ok}(s_0)$ is the ordinary kriging predictor of $S(s_0)$. Observe that the ordinary kriging predictor of $X(s_0)$ is biased, but since $X(\cdot)$ is a log Gaussian process, a bias correction can

be applied to the predictor and the MSPE of the corrected predictor is easy to obtain (e.g., Cressie, 1993a; Sect. 3.2.2).

Suppose that $E(S(s)) \equiv \mu$, $\text{var}(S(s)) \equiv \sigma^2$, and $\text{var}(\mathbf{Z}) \equiv \Sigma$. Then, adjusting for measurement error, the bias-corrected ordinary kriging predictor of $X(s_0)$ is given as follows,

$$\hat{X}_{ok}(s_0) \equiv \exp\{\hat{S}_{ok}(s_0) + (\sigma^2 - \text{var}(\hat{S}_{ok}(s_0)))/2\},$$

and the MSPE of $\hat{X}_{ok}(s_0)$ is given by

$$\begin{aligned} \text{MSPE}(\hat{X}_{ok}(s_0)) &\equiv \{\exp(2\mu + \sigma^2)\} \\ &\times \{\exp(\sigma^2) + \exp(\text{var}(\hat{S}_{ok}(s_0))) - 2\exp(\text{cov}(S(s_0), \hat{S}_{ok}(s_0)))\}. \end{aligned}$$

Define the constrained kriging predictor of $X(s_0)$ as

$$\hat{X}_{ck}(s_0) \equiv \exp(\hat{S}_{ck}(s_0)); \quad s_0 \in D,$$

where $\hat{S}_{ck}(s_0)$ is the constrained kriging predictor of $S(s_0)$. Observe that since $\exp(\cdot)$ is a smooth function, the constrained kriging predictor of $X(s_0)$ is at least approximately unbiased, whatever the distribution of $S(\cdot)$. However, since we have assumed that $S(\cdot)$ is Gaussian, the constrained kriging predictor $\hat{X}_{ck}(s_0)$ is exactly unbiased. It is not difficult to show that its MSPE is given by

$$\text{MSPE}(\hat{X}_{ck}(s_0)) \equiv 2\{\exp(2\mu + \sigma^2)\}\{\exp(\sigma^2) - \exp(\text{cov}(S(s_0), \hat{S}_{ck}(s_0)))\}.$$

Contour maps of the bias-corrected ordinary kriging predictions and the constrained kriging predictions of PM_{10} over D are presented in Figure 7.4, and contour maps of their respective MSPEs are presented in Figure 7.5. Notice that the map of ordinary kriging predictions appears to be a little smoother than that of constrained kriging predictions, and that the MSPEs of the ordinary kriging predictions are quite a bit larger than those of constrained kriging.

Inspection of Figure 7.4 shows that the region D is well within PM_{10} compliance on day 226, 1996, since the largest predicted value is well below the 24-hour standard of $150 \mu\text{g}/\text{m}^3$. However, for illustrative purposes, suppose that we wish to know the proportion of D for

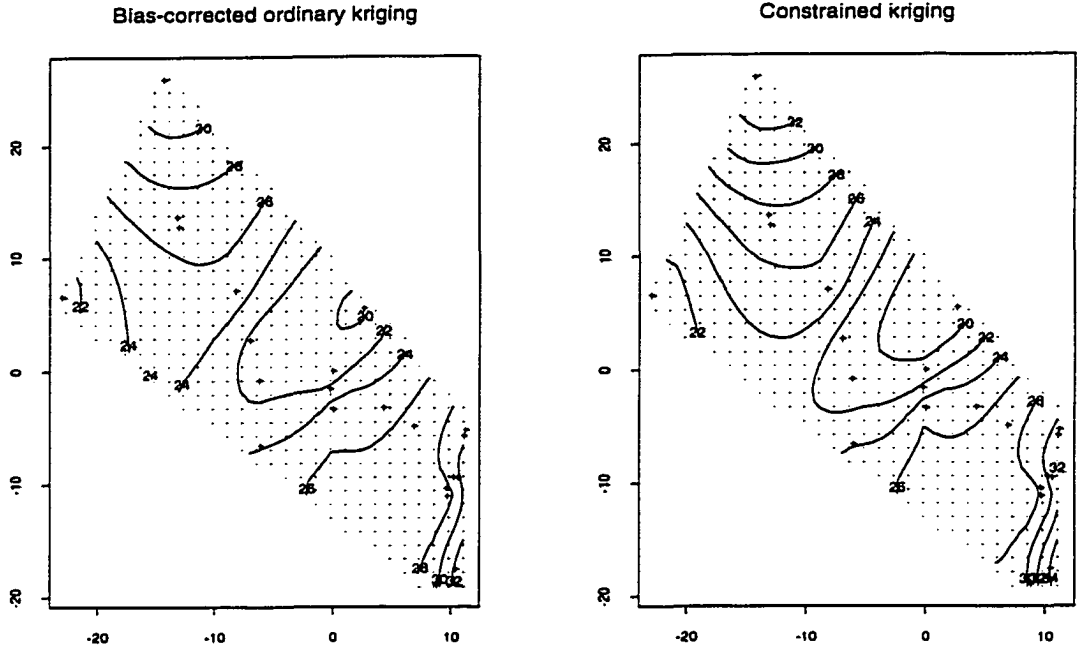


Figure 7.4 Contour maps of bias-corrected ordinary kriging and constrained kriging predictions of PM_{10} . There are 500 prediction locations marked “.”, and 22 sampling locations marked “+”.

which PM_{10} exceeds the two cutoff values of 25 and $30 \mu g/m^3$. The predictand describing this quantity is defined as

$$\begin{aligned} F_X(t) &\equiv \frac{1}{|D|} \sum_{s \in D} I(X(s) \leq t) \\ &= \frac{1}{|D|} \sum_{s \in D} I(\exp(S(s)) \leq t); \quad t = 25, 30. \end{aligned}$$

Further, suppose we wish to predict the average amount of PM_{10} that exceeds/does not exceed the two cutoff values. The predictands describing these quantities can be written respectively as

$$\begin{aligned} Q_A(t) &\equiv \frac{1}{|D|} \sum_{s \in D} X(s) I(X(s) > t) / (1 - F_X(t)) \\ &= \frac{1}{|D|} \sum_{s \in D} \exp(S(s)) I(\exp(S(s)) > t) / (1 - F_X(t)); \quad t = 25, 30, \end{aligned}$$

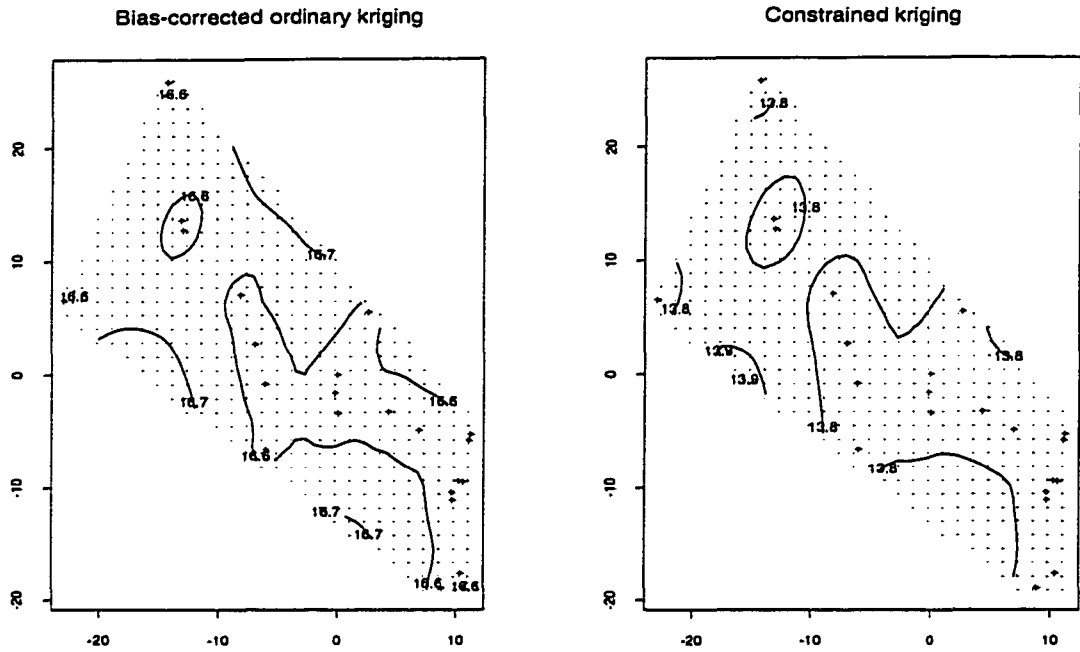


Figure 7.5 Contour maps of bias-corrected ordinary kriging and constrained kriging MSPEs of PM_{10} predictions. There are 500 prediction locations marked “.”, and 22 sampling locations marked “+”.

and

$$\begin{aligned}
 Q_A^c(t) &\equiv \frac{1}{|D|} \sum_{s \in D} X(s) I(X(s) \leq t) / F_X(t) \\
 &= \frac{1}{|D|} \sum_{s \in D} \exp(S(s)) I(\exp(S(s)) \leq t) / F_X(t); \quad t = 25, 30.
 \end{aligned}$$

Ordinary kriging, constrained kriging, and covariance-matching constrained kriging were used to predict the predictands given above. The covariance-matching constrained kriging predictor was constructed as in Section 6, that is, the 500×1 prediction vector S was partitioned into subvectors of length 5 (i.e., $S \equiv [S(s_1), \dots, S(s_5) | S(s_6), \dots, S(s_{10}) | \dots | S(s_{496}), \dots, S(s_{500})]'$, where the numbering of the prediction locations proceeds sequentially from left to right, starting at the bottom row and ending at the top row). The matrices P_o and Q_o were decomposed into symmetric components according to the spectral decomposition method. The

prediction results given in Table 7.7 are fairly similar for the three predictors, except in predicting $F_X(25)$, where the covariance-matching constrained kriging prediction is substantially different from the other predictions, and the ordinary kriging prediction of $F_X(30)$ is somewhat larger than the others.

Table 7.7 Ordinary kriging, constrained kriging, and covariance-matching constrained kriging predictors of $F_X(t)$, $Q_A(t)$, and $Q_A^c(t)$; $t = 25, 30$. Ordinary kriging is denoted "OK", constrained kriging "CK", and covariance-matching constrained kriging "CM".

	$F_X(t)$		$Q_A(t)$		$Q_A^c(t)$	
	$t = 25$	$t = 30$	$t = 25$	$t = 30$	$t = 25$	$t = 30$
OK	0.466	0.942	27.30	31.13	22.86	24.87
CK	0.472	0.906	27.95	32.09	22.21	24.53
CM	0.360	0.908	27.82	33.21	22.96	25.35

8 Conclusions

Several conclusions can be drawn from this study. The first is that if a measurement-error component is explicitly added to a spatial model, then this may have a profound effect on the nature of spatial prediction. In particular, we show that nonlinear kriging predictors do not easily accommodate measurement error, and these predictors are shown to be biased in the presence of measurement error. On the other hand, if we assume an additive measurement-error component in a spatial model, then the linear kriging predictors easily filter out the measurement error. The unbiasedness of the linear predictors remains unchanged in the presence of measurement error and the mean-squared prediction error simply includes the measurement-error variance as an additive component. In addition, the spatial model with additive measurement error can handle replicated data (i.e., multiple observations at a single sampling location) and data where the same underlying spatial phenomenon is observed via two or more distinct types of measuring instrument.

The second conclusion is that constrained kriging appears to outperform ordinary kriging in the prediction of quantities such as $g(S(B))$, where g is some nonlinear function, particularly when $S(\cdot)$ is Gaussian. However, the covariance-matching constrained kriging predictor, or

some “partitioned” version of it (see (5.24)), generally appears to outperform constrained kriging for any measurable functional of $S(\cdot)$, if $S(\cdot)$ is Gaussian.

Ideas for future research include the role of conditional simulation (based on simple kriging) as a possible alternative to constrained kriging in cases where constrained kriging does not exist, and the effect of the decomposition of \mathbf{P}_u and \mathbf{Q}_u (see Section 5.1.3) on the optimality of the covariance-matching constrained kriging predictor.

References

- Aldworth, J. and Cressie, N. (1998), “Sampling designs and prediction methods for gaussian spatial processes”, in *Multivariate Analysis, Design of Experiments and Survey Sampling*, S. Ghosh (ed.), Marcel Dekker, New York, forthcoming.
- Cressie, N. (1985), “Fitting variogram models by weighted least squares”, *Journal of the International Association of Mathematical Geology*, 17, 563–586.
- Cressie, N. (1986), “Kriging nonstationary data”, *Journal of the American Statistical Association*, 81, 625–634.
- Cressie, N. (1988), “Spatial prediction and ordinary kriging”, *Mathematical Geology*, 20, 405–421 [Erratum: *Mathematical Geology*, 21, 493–494.].
- Cressie, N. (1993a), *Statistics for Spatial Data*, rev. edn., Wiley, New York.
- Cressie, N. (1993b), “Aggregation in geostatistical problems”, in *Geostatistics Tróia 1992*, Vol. 1, A. Soares (ed.), Kluwer Academic Publishers, Dordrecht, pp. 25–36.
- Cressie, N., Kaiser, M. S., Daniels, M. J., Aldworth, J., Lee, J., Lahiri, S. N., and Cox, L. H. (1999), “Spatial analysis of particulate matter in an urban environment”, in *geoENV2 — Geostatistics for Environmental Applications*, A. Soares, J. Gomez-Hernandez, and R. Froidevaux (eds.), Kluwer Academic Publishers, Dordrecht, forthcoming.
- Fuller, W. A. (1995), “Estimation in the presence of measurement error”, *International Statistical Review*, 6, 121–147.
- Gotway, C. A. and Cressie, N. (1993), “Improved multivariate prediction under a general linear model”, *Journal of Multivariate Analysis*, 45, 56–72.
- Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

- Horvitz, D. G. and Thompson, D. F., (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, **47**, 663–685.
- Journal, A. G. (1974), "Geostatistics for conditional simulation of ore bodies", *Economic Geology*, **69**, 673–687.
- Journal, A. G. (1983), "Nonparametric estimation of spatial distributions", *Journal of the International Association of Mathematical Geology*, **15**, 445–468.
- Journal, A. G. and Huijbregts, C. J. (1978), *Mining Geostatistics*, Academic Press, London.
- Lahiri, S. N., Kaiser, M. S., Cressie, N., and Hsu, N.-J. (1999), "Prediction of spatial cumulative distribution functions using subsampling", *Journal of the American Statistical Association*, forthcoming.
- Lajaunie, C. (1990), "Comparing some approximate methods for building local confidence intervals for predicting regionalized variables", *Mathematical Geology*, **22**, 123–144.
- Lantuejoul, C. (1988), "On the importance of choosing a change of support model for global reserves estimation", *Mathematical Geology*, **20**, 1001–1019.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.
- Matheron, G. (1963), "Principles of geostatistics", *Economic Geology*, **58**, 1246–1266.
- Myers, D. E. (1982), "Matrix formulation of co-kriging", *Journal of the International Association of Mathematical Geology*, **14**, 249–257.
- Overton, W. S. (1993), "Probability sampling and population inference in monitoring programs", in *Environmental Modeling with GIS*, M. F. Goodchild, B. O. Parks, and C. T. Steyaert (eds.), Oxford University Press, New York, pp. 470–480.
- Rao, C. R. (1985), "Matrix derivatives: Applications in statistics", in *Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz and N. L. Johnson (eds.), Wiley, New York.
- Rivoirard, J. (1994), *Introduction to Disjunctive Kriging and Non-linear Geostatistics*, Oxford University Press, Oxford.
- Särndal C.-E., Swensson B., and Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Schott, J. R. (1997), *Matrix Analysis for Statistics*, Wiley, New York.

- Stefanski, L. A. and Bay, J. M. (1996), "Simulation extrapolation deconvolution of finite population cumulative distribution function estimators", *Biometrika*, **83**, 406–417.
- U. S. Environmental Protection Agency, (1996a), "Air quality criteria for particulate matter", Final Draft, National Center for Environmental Assessment, Office of Research and Development, U. S. EPA, Research Triangle Park, NC.
- U. S. Environmental Protection Agency, (1996b), "Review of national assessment of scientific and technical information", OAQPS Staff Paper, EPA report no. 452/R-96-013, Office of Air Quality Planning and Standards, U. S. EPA, Research Triangle Park, NC.
- Ver Hoef, J. M. and Cressie, N. (1993), "Multivariable spatial prediction", *Mathematical Geology*, **25**, 219–240 [Errata: *Mathematical Geology*, **26**, 273–275.].

Acknowledgements

The research presented in this article was supported by the U. S. Environmental Protection Agency under co-operative agreement #CR822919-01-0. The article has not been subjected to the review of the EPA and thus does not reflect the view of the agency and no official endorsement should be inferred.

GENERAL CONCLUSIONS

In this dissertation, we have examined the role of measurement error in spatial models, the prediction of nonlinear functionals of spatial processes, and spatial-sampling theory.

In the first paper, several conclusions can be drawn, but it should be noted that these conclusions pertain to the characteristics of the spatial phenomenon described in Section 3.1 of this paper. In particular, this spatial phenomenon does not contain values of interest that are either clustered or rare.

Designs that correspond to so-called “representative-site” selection should be avoided. The choice of sampling design from among systematic, stratified-random, or simple-random sampling designs appears to be unimportant for both spatial-mean and SCDF prediction.

For spatial-mean prediction (i.e., a linear predictand) over the local region, the spatial BLUP (i.e., ordinary kriging) is the preferred predictor, although constrained kriging performs competitively, especially for stationary processes. Both predictors require that the spatial covariance parameters be known or well estimated. The regional poststratification predictor should be avoided if the measurement error is large, and the arithmetic mean should be avoided in the presence of a trend component.

For SCDF prediction (i.e., a nonlinear predictand) over the global region, the so-called “best predictor” performs best, but requires the strongest model assumptions. Constrained kriging performs well and requires fewer model assumptions. The simplified-model, deconvolution, and Horvitz-Thompson predictors perform well only if the measurement-error component is small. Ordinary kriging should be avoided.

Effects of different factors/levels on SCDF prediction are only discernible for larger sample sizes, in comparison to those for spatial-mean prediction. In those cases, constrained kriging is a superior predictor.

The conclusions stated above were generally consistent across the three S -realizations generated, with minor exceptions as noted in Section 4 of the first paper.

In the second paper, several conclusions also can be drawn. The first is that if a measurement-error component is explicitly added to a spatial model, then this may have a profound effect on the nature of spatial prediction. In particular, we show that nonlinear kriging predictors do not easily accommodate measurement error, and these predictors are shown to be biased in the presence of measurement error. On the other hand, if we assume an additive measurement-error component in a spatial model, then the linear kriging predictors easily filter out the measurement error. The unbiasedness of the linear predictors remains unchanged in the presence of measurement error and the mean-squared prediction error simply includes the measurement-error variance as an additive component. In addition, the spatial model with additive measurement error can handle replicated data (i.e., multiple observations at a single sampling location) and data where the same underlying spatial phenomenon is observed via two or more distinct types of measuring instrument.

The second conclusion is that constrained kriging appears to outperform ordinary kriging in the prediction of quantities such as $g(S(B))$, where g is some nonlinear function, particularly when $S(\cdot)$ is Gaussian. However, the covariance-matching constrained kriging predictor, or some "partitioned" version of it (see (5.24) in the second paper), generally appears to outperform constrained kriging for any measurable functional of $S(\cdot)$, if $S(\cdot)$ is Gaussian.

Ideas for future research include the role of conditional simulation (based on simple kriging) as a possible alternative to constrained kriging in cases where constrained kriging does not exist, and the effect of the decomposition of \mathbf{P}_u and \mathbf{Q}_u (see Section 5.1.3 in the second paper) on the optimality of the covariance-matching constrained kriging predictor.

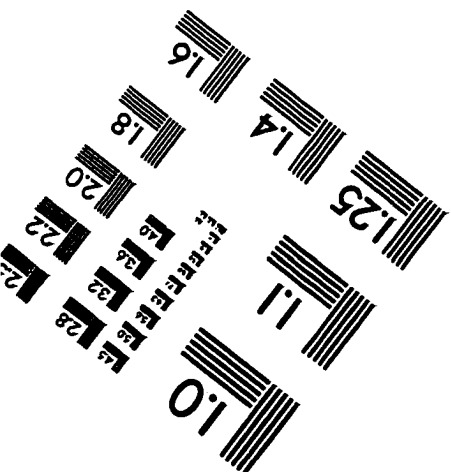
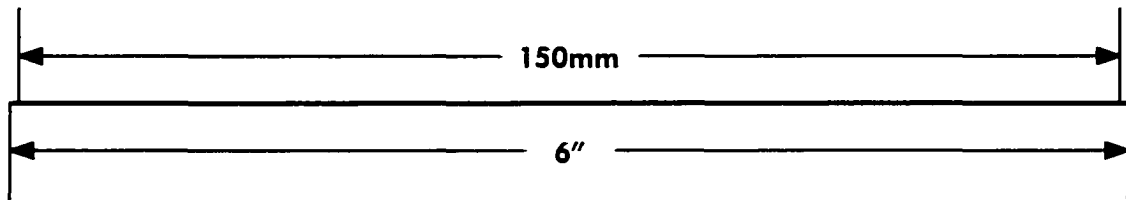
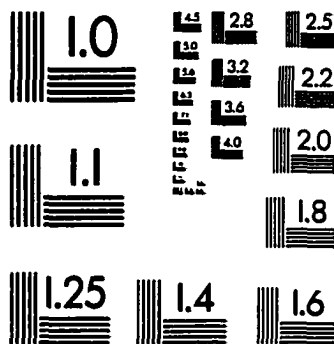
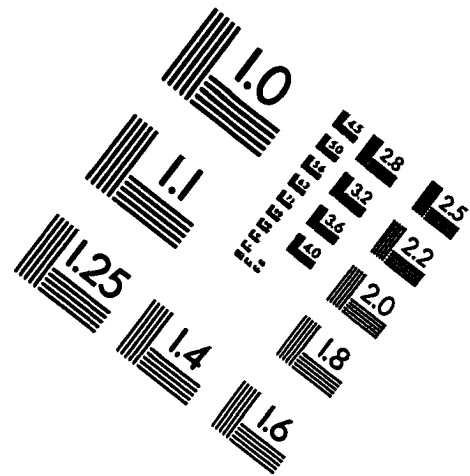
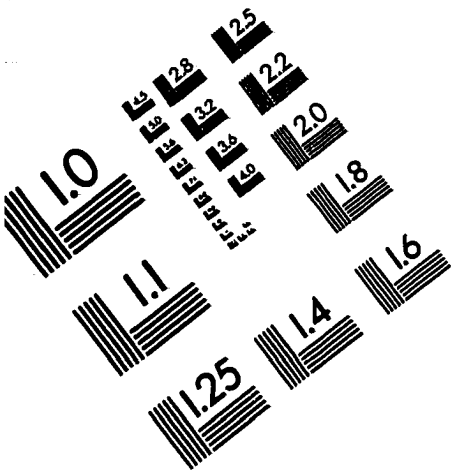
ACKNOWLEDGEMENTS

The research presented in this dissertation was supported by the U. S. Environmental Protection Agency under co-operative agreement #CR822919-01-0.

The people who have made this work possible are many and go back a long way. Perhaps I should start with Brother Adrian, a contemplative Irishman from Dingle, Co. Kerry, from whom I first learnt to appreciate the proportion and elegance in the “maths” that he taught at high school. The list goes on, with too many names to mention individually, of the many excellent professors of mathematics and statistics who also have inspired me through the years. Above all, I would like to express my appreciation to Professor Noel Cressie, from whom I learnt a tremendous amount.

Finally, I dedicate this work to my wife, Sue, who always believed I could do it. As a token of my appreciation, I do not expect her to read this dissertation!

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved

